Proyecto de Sistemas Informáticos
Facultad de Informática
Universidad Complutense de Madrid

Evaluación y selección de software. Extracción automática de texto en ficheros PDF.

Enrique Fernández Martínez

Alberto López Rubio

Profesora directora: Mónica Chagoyen Quiles

Palabras para búsqueda bibliográfica:

consultoría, extracción, conversión, automática, ficheros, texto

PDF, TXT y XPDF.

Los alumnos que hemos desarrollado este proyecto autorizamos a la Universidad Complutense de Madrid a difundir y a utilizar con fines académicos, no comerciales y mencionando expresamente a sus autores, el informe aquí presentado.

Enrique Fernández Martínez

Alberto López Rubio

ÍNDICE

Resumen

Se trata de un proyecto de consultoría/evaluación tecnológica con el objeto de definir un sistema para la extracción "estructurada" del texto de artículos científicos (concretamente en el área de biomedicina) almacenados en formato PDF.

PubMed Central (www.pubmedcentral.nih.gov) es el archivo digital del NIH estadounidense (National Institutes of Health) que ofrece acceso a los artículos publicados en las áreas de biomedicina y ciencias de la vida . PubMed Central ha definido un estandar (en formato DTD) para la estructuración del contenido de dichos artículos. Sin embargo el uso de este estándar no está muy extendido y la mayoría de publicaciones se encuentran en formato PDF. La cantidad de publicaciones hoy en día en tan grande que hace casi imposible encontrar la información que pueda estar relacionada con un proyecto sin un formato estándar. De ahí nace la iniciativa del NIH y la necesidad de una herramienta para convertir documentos en otros formatos a un mismo estándar.

En la actualidad existen diversos programas para la conversión de archivos en formato PDF a texto, el objetivo del proyecto es determinar cual sería el software más adecuado para esta conversión. En un primer análisis se han realizado varias pruebas con algunos de los programas más destacados con distinto tipo de licencias. Entre ellos se han elegido los mejores y se ha hecho un análisis más exhaustivo comprobando todas las funcionalidades de cada uno de ellos.

Cabe destacar el XPDF cuyo código fuente está disponible bajo licencia GNU y que permitiría trabajar sobre él para una futura adaptación al formato NCBI DTD.

Abstract

The project is about the technology consulting/evaluation for the definition of a “structured” extraction system of text for scientific publications(more exactly in the biomedicine area) stored in PDF format.

PubMed Central (www.pubmedcentral.nih.gov) is the U.S. National Institutes of Health (NIH) digital archive of biomedical and life sciences journal literature. PubMed Central has established a standard (in DTD format) for the organization of article contents. However the use of this standard is not very spread and nowadays most of publications are in PDF format. The quantity of publications is so huge that make it impossible to find the related information to a project without a standard format. This was the reason for the NIH initiative and the need of a tool to convert documents in other formats to a common standard.

Nowadays there are different programs for the conversion of PDF formatted files to text files, the goal of this project is to determine wich would be the most relevant software to make this convertion. In a first analysis we have done different tests with some of the best programs with different licence type. Among them we have chosen the bests and make a more detailed test checking all available functions.

It is worth pointing out XPDF that its source code is available under the GNU licence so it would allow to work on it for a future adaptation to the NCBI DTD format.

1. INTRODUCCIÓN

1.1 OBJETIVO DEL USUARIO

El usuario desea desarrollar una solución que permita un procesamiento automático posterior (mediante técnicas de minería de texto y extracción de la información) del contenido textual de artículos científicos publicados en el área de la Biomedicina. Dichos artículos son generalmente accesibles 'on-line' en dos formatos alternativos, HTML y PDF, mientras que las herramientas de procesamiento automático del usuario trabajan únicamente con texto plano.
Por diversas razones técnicas y de disponibilidad el usuario desea procesar únicamente ficheros PDF. Se desea además, disponer del contenido de los artículos de manera estructurada (se propone utilizar como modelo de estructura el DTD elaborado por el NCBI para los artículos almacenados en PubMed Central).

El desarrollo de dicha solución se realizará en dos etapas:

Etapa primera (consultoría técnica): identificación de la herramienta para realizar la conversión de ficheros PDF a texto

Etapa segunda (desarrollo): elaboración de la(s) herramienta(s) que realicen la estructuración del contenido de los artículos

El presente proyecto abarca la primera etapa de consultoría técnica, por lo tanto, el objetivo concreto del proyecto es:

Evaluar el software disponible para la conversión de documentos PDF a texto, seleccionando aquella solución que mejor cubra los requerimientos del usuario/cliente.

1.2 ESTRATEGIA DE TRABAJO

Se trata del estudio y clasificación de algunas de las herramientas software actuales para la extracción automática de texto en los ficheros pdf, especialmente del extractor de texto XPDF.

Inicialmente se realizará un estudio superficial de algunas de estas herramientas tales como presentación de dicha herramienta, funciones que nos posibilita realizar, entorno de trabajo y otros factores que se puedan tener en cuenta.

Después de este estudio previo se procederá a una comparativa entre todas las diversas herramientas, se intentara analizar tanto las características comunes, como las que las hacen diferentes una de las otras, en este estudio se realizara un descarte definitivo de 6 de las herramientas para al finar quedarnos con 2 de ellas que serán las que se estudien a fondo en el último de los estudios.
Este descarte se realizara en función a lo completa y correctamente funcional que la aplicación sea.

El tercero de los estudios ya solo sobre 2 de las herramientas consistirá en una evaluación de rendimiento tanto a nivel de consumo de recursos de CPU, como de consumo de memoria RAM de las herramientas que han sido seleccionadas, también se realizará un estudio de productividad de dichas herramientas, dicho estudio queda justificado ya que también es necesario tener en cuenta un gran factor y es que queremos que esta aplicación se use en una empresa y esta necesita productividad a parte de una buena calidad en sus productos.

Cada uno de estos estudios estará estructurado de la siguiente manera:

1 Consideraciones Iníciales: Como se realizara el estudio.

2 Resultados del estudio: Los resultados obtenidos.

3 Conclusiones del estudio: Realizadas a partir del estudio de los resultados.

En las conclusiones del último estudio ya se procederá al descarte definitivo y se especificara que herramienta es la que ha sido seleccionada como la mejor, entre las que han sido analizadas.

2. EXPERIMENTACION

2.1 PRIMER ESTUDIO

2.1.1 CONSIDERACIONES INÍCIALES

ASPECTOS TENIDOS EN CUENTA DURANTE LA PRIMERA EVALUACIÓN:

En esta primera prueba se tendrán en cuenta los aspectos más generales de cada uno de estas herramientas desde el formato de presentación, el precio de dicho programa, el uso y el fichero de salida obtenido, el comportamiento y otras múltiples cosas que se podrán ver a continuación.

BANCO DE TRABAJO.

PROGRAMAS QUE SERÁN ANALIZADOS:

- **SHAREWARE Y FREEWARE**
  - XPDF.
  - Cool PDF READER
  - Easy PDF to Text converter
  - A-PDF Text Extractor
- **COMERCIAL**
- PDF TEXT Converter
- LD GETTER
- LD GETTER PRO
- PDF2TEXT
- PDF Plain Text Extractor

FICHERO DE ENTRADA SOBRE EL CUAL SE REALIZARA LA PRUEBA:

- http://www.biomedcentral.com/content/pdf/1471-213X-5-14.pdf

Programa:	XPDF (Programa referencia)
Donde encontrarlo:	http://www.foolabs.com/xpdf
Tipo:	Open Source
Licencia:	GNU
Presentación:	Se puede obtener tanto el código fuente del programa y compilarlo tú usando tus propias librerías o descargarse una versión previamente compilada. Nosotros hemos optado por la descarga de una versión ya compilada exactamente la versión: Win32 (built with MSVC, esta versión no incluye el visor de pdfs, ya que el visor solo es para entornos gráficos X. Esta versión solo incorpora las herramientas de extracción de texto, imágenes, meta-información y fuentes de los ficheros de texto, lo cual no es un problema puesto que lo único que necesitamos es el extractor de texto ya que es nuestro objeto de estudio. La versión en cuestión es: xpdf-3.02pl2-win32.zip (2027995 bytes) que como se ve se trata de un fichero zip.
Instalación:	Como única opción de instalación que encontramos es la descompresión de dicho fichero, puesto que en su interior ya se encuentran los ficheros ejecutables y los ficheros de ayuda de las diversas herramientas incluidas en el fichero zip.
Utilización:	De los diversos ficheros dentro del archivo zip, nos centraremos en dos. El primero de ellos es el fichero de ayuda pdftotext.txt, este fichero contiene la información de versión, autor, integración y una breve descripción de los diversos flags que se pueden activar al inicio de la extracción de texto, flags como si queremos el texto en formato html con la meta-informacion del fichero pdf, también podemos indicarle el número de páginas que se quieren convertir, si queremos que intente mantener la distribución original del texto… También incluye los códigos de salida, indicando si se produce algún error durante la conversión del texto. El segundo de ellos es la herramienta en sí, se trata de un fichero ejecutable que hay que hacer correr sobre línea de comandos, la utilización es la estándar de este tipo de programas, es decir pdftotext ficherodeseado.pdf y los diversos flags que queramos usar, si se especifica nombre de fichero de salida se usara el especificado sino se usara el propio del fichero de entrada con extensión.txt
Resultado obtenido:	Para mayor sencillez en la comprobación de resultados se ha solicitado al programa que conserve el diseño de entrada, pero por defecto genera cadenas de caracteres con la longitud total del párrafo procesado. Comparando el fichero de entrada con el fichero de salida se obtiene que extrae todo el texto sin distinguir si son marcas de pagina o encabezados, estos encabezados en la versión sin layout aparecen en dos líneas situando como primera la línea la parte más a la izquierda y como segunda la parte más a la derecha del encabezado, en la versión con layout aparece en una sola línea lo cual será más sencillo de para un preprocesado posterior, además dichas líneas aparecen con una marca o, de carácter no imprimible; en cuanto aparecen imágenes en el texto y dichas imágenes van acompañadas de un texto aclaratorio, el titulo del texto aparece dos veces, si se usa layout intenta dejar el hueco de la imagen mediante el uso de líneas en blanco, las referencias en la versión con layout son fácilmente obtenibles, pero en la versión sin layout puede ser laborioso puesto que se obtiene por un lado la numeración y a continuación la referencia en sí misma, por lo que asociar la referencia con lo referido puede resultar complicado.

Programa:	Cool PDF READER
Donde encontrarlo:	http://www.pdf2exe.com
Tipo:	FreeWare
Licencia:	FreeWare
Presentación:	Encontrar la herramienta en la pagina cuesta un poco puesto que no aparece un extractor como era de esperar, la herramienta que buscamos es la que denominan Reader. En cuanto a la descarga se hace de manera rápida y sencilla. De las tres distribuciones que nos presentan todas ellas para entornos Windows hemos escogido la versión Portable o como ellos la denominan la StanAlone versión. Se trata de esta versión: Standalone Package with no installation required (just unzip & run) 626KB
Instalación:	Al tratarse de una distribución portable lo único que tienes que hacer es descargar y ejecutar no requiere ni instalación siquiera.
Utilización:	Al hacer doble click sobre el ejecutable se nos presenta una interfaz grafica con botones intuitivos tipo Office 2003, lo cual facilita el reconocer los diversos tipos de funcionalidad que presenta la aplicación la única ayuda que presenta el programa son los ToolTips de los botones y un enlace a la página de ayuda. La forma de uso es la siguiente cargas el fichero en el programa, te muestra la primera pagina del pdf, y entonces decides intentar extraer el texto, la única forma de hacerlo es mediante la opción guardar, además dicha opción solo te permite guardar//convertir la pagina actual, si el fichero es un poco extenso la tarea puede ser tediosa.
Resultado obtenido:	Como así nos lo brinda el programa obtenemos la primera pagina, donde se encuentra el Abstract (Parte de interés del estudio), la primera diferencia que nos encontramos frente al programa referencia, es que directamente preserva la distribución del texto aunque nosotros no se lo solicitemos, pero la distribución la hace de una manera un tanto curiosa, primero el titulo del texto, y luego la pagina en la que estamos (esto lo ha reubicado puesto que viene al final de la pagina y esta al principio) y luego ya se encuentra el texto en una distribución normal, también incorpora la cabecera al igual que el programa de referencia, al encontrarse con imágenes en el texto, opta por primero incorporar todo el texto y al final del texto extraído es donde aparecen las referencias de las imágenes y los cuadros de texto existentes. En la parte de las referencias vuelve a reorganizar el texto primero coloca lo que él considera texto plano, es decir las referencias en la primera parte y todo lo que se encuentra en cuadros de texto, lo coloca en la parte posterior, las referencias quedan claras y visibles.
Conclusiones:	El programa presenta una interfaz grafica agradable pero no es lo suficientemente clara, la extracción única de la pagina actual es algo que si el volumen de trabajo es muy alto, puede resultar algo incomodo, y si se quiere realizar un posterior trabajo con todo el contenido del fichero, nos vemos en la obligación de tener que realizar esta tarea de forma manual. Los ficheros individuales obtenidos son altamente legibles y bastante similares al de entrada con las salvedades comentadas anteriormente, las imágenes que encuentra en el texto las extrae como ficheros jpg de manera automática siempre que puede.

Programa:	Easy PDF to Text converter
Donde encontrarlo:	http://www.pdf-to-html-word.com/pdf-to-text/
Tipo:	FreeWare
Licencia:	FreeWare
Presentación:	Solo se encuentra una única versión que haga lo que nosotros estamos buscando, dicha distribución solo está disponible para entornos Windows. Este programa al igual que los dos anteriores no requiere librerías específicas de adobe, usa las suyas propias para la extracción. La versión en cuestión es Easy PDF to Text Converter v2.0.0 Installation Wizard.
Instalación:	Presenta instalación tipo asistente de Windows, se instala sin ningún problema y las opciones de configuración son inexistentes únicamente nos permite escoger si queremos crear iconos en el escritorio.
Utilización:	Al tratarse de una instalación de asistente típica el programa lo encontraremos en el grupo de trabajo que nosotros hayamos escogido, también se encuentra la ayuda del programa en dicha carpeta, al solicitarla obtendremos el clásico visor de ayuda de Windows. Una vez que lancemos el programa nos encontramos una interfaz grafica, sencilla y funcional cargamos el fichero de entrada dándole al botón al uso y en el segundo cuadro dialogo le damos el nombre al fichero de salida, durante la conversión sensiblemente más lenta que el resto podemos ver una barra de progreso que se va rellenando según procesa las diversas paginas.
Resultado obtenido:	Al finalizar la conversión podemos encontrar en la carpeta de salida un fichero de texto plano por cada una de las paginas que tenía el fichero pdf de origen, el nombre de los archivos se corresponde al escogido por nosotros concatenado con el numero de página de la que se trata. Se trata de una conversión bastante mala, ya que intenta convertir todo el texto que encuentra sea como sea, convierte por líneas mezclando las columnas que hay en el texto lo cual hace que se complicado de seguir en la conversión, cuando encuentra negritas o cursivas en el texto original o a veces sin incluso encontrarlas coloca todo el texto de una manera continuada en el fichero de manera que no se sabe lo que se está leyendo. En cuanto a la aparición de imágenes, no las extrae automáticamente y coloca el texto donde deberían estar, también replica la información de titulo de las imágenes. La pagina que contiene las referencias resulta ilegible al extraer el texto por líneas mezcla cosas de ambos lados por lo cual luego es imposible realizar la lectura del texto o un posterior procesado.
Conclusiones:	Aunque sea el programa más sencillo e intuitivo de todos, la conversión que realiza es bastante mala, si el fichero de entrada fuese un fichero plano sin demasiado formato, podría ser una buena opción para trabajos sencillos pero no es nada recomendable posiblemente sea una de las peores opciones analizadas.

Programa:	A-PDF Text Extractor
Donde encontrarlo:	http://www.a-pdf.com/text/
Tipo:	FreeWare
Licencia:	FreeWare (Admiten donaciones)
Presentación:	Se presentan dos versiones ambas para Windows una versión con entorno grafico que es la que estudiaremos aquí puesto que es la que es versión freeware, mientras que la versión de línea de comandos es de pago y por lo tanto no accesible en este momento. La versión en si es: A-PDF Text Extractor v1.1.0
Instalación:	Presenta instalación tipo asistente de Windows, se instala sin ningún problema y las opciones de configuración son inexistentes únicamente nos permite escoger si queremos crear iconos en el escritorio.
Utilización:	Al tratarse de una instalación de asistente típica el programa lo encontraremos en el grupo de trabajo que nosotros hayamos escogido, también se encuentra una página HTML llamada How To Use, que nos cuenta cómo usar el programa en sí. Una vez que lancemos el programa nos encontramos una interfaz grafica, sencilla y funcional cargamos el fichero de entrada dándole al botón al cargar el fichero nos informa sobre el numero de páginas de el mismo y nos invita a pulsar el botón de extracción de texto, tenemos un cuadro de opciones que nos permite seleccionar si queremos extraer todas las paginas o solo un rango, si queremos las partes impares o todas, o si queremos incorporar un cabecera y un pie de página. Le damos a extraer seleccionamos el fichero de salida realiza la conversión rápidamente.
Resultado obtenido:	Este programa también reordena el texto convertido colocando el texto de la numeración de pagina debajo del título del artículo y al comienzo de cada una de las paginas, al final de cada página incorpora una marca de fin de pagina sencilla una cosa así: = Page 1 = Extrae tanto la información de los pies de pagina como de las cabeceras, lo que considera que se encuentra en cuadros de texto no referentes al texto central los coloca también al final de la conversión, respeta correctamente las columnas y lo coloca de forma que es un renglón de la columna una línea en el fichero de salida. Al encontrar imágenes lo que hace es colocar el texto referente a estas en la parte baja de la página y replica el titulo de las fotos como el resto hace. En la parte de las referencias, estas se encuentran al principio de la página y todo lo que se encuentra en cuadros de texto se puede ver al final de la página, estas referencias son claras y bien visibles y fácilmente procesables.
Conclusiones:	Sencillo rápido y muy claro en la conversión, también intenta convertir los títulos de los encabezados de las paginas, hasta ahora el único que lo hace, el marcado que hace con los ficheros de salida de las paginas hace muy sencillo en un procesado posterior el descarte de páginas enteras, lo cual puede venir muy bien para posteriores procesamientos. En resumen es una buena herramienta.

Programa:	PDF2TEXT
Donde encontrarlo:	http://www.pdftoall.com/
Tipo:	Profesional
Licencia:	ShareWare (La licencia cuesta 71,85€)
Presentación:	Se presenta una única versión para Windows, además es necesario tener instalado el Adobe Acrobat puesto que utiliza las librerías de adobe para ser instalado, usaremos la única versión disponible en la web.
Instalación:	Se presenta con un fichero zip donde podemos encontrar el fichero de instalación que es el típico asistente de Windows, se instala sin ningún problema y las opciones de configuración son inexistentes únicamente nos permite escoger si queremos crear iconos en el escritorio.
Utilización:	Al tratarse de una instalación de asistente típica el programa lo encontraremos en el grupo de trabajo que nosotros hayamos escogido, también se encuentra la ayuda, la dirección web de contacto y el acuerdo de licencia que hemos escogido en este caso Trial. Una vez que lancemos el programa nos encontramos una interfaz grafica, sobria y funcional,el programa permite por ejemplo la conversión en serie de diversos ficheros pdf, cargamos los diversos ficheros que queramos convertir en este caso nuestro fichero de prueba, mediante la opción add, en las opciones le podemos especificar ante lo que se encuentra si es un documento o si es un informe (Hemos marcado informe y se come la mitad de las palabras) y también hay que especificarle si el fichero de entrada tiene dos columnas.
Resultado obtenido:	No podemos realizar un análisis tan extenso como en los programas de versión libre puesto que al ser una versión ***trial***, solo nos permite el extraer las paginas pares, además es bastante lento en comparación con los demás y justo al finalizar la conversión dispara el consumo de CPU de una manera espectacular. Extrae también los encabezados y los pies de página y en presencia de imágenes, coloca el texto donde estas deberían ir, replicando la información del título de las imágenes. Al haber seleccionado que el texto está en dos columnas, también usa el método de un renglón una línea lo cual hace bastante sencillo el seguir el texto, si no se hubiese marcado esa opción el resultado dejaba bastante que desear. En cuanto a la página de referencias coloca la información de los cuadros de texto por encima del texto plano y las referencias quedan claramente visibles y sencillos de leer.
Conclusiones:	Entorno profesional y de pago, permite la conversión de múltiples ficheros en serie lo ponen como algo bueno, pero realmente no lo es ya que para que sea realmente eficaz los ficheros deben de haber sido previamente filtrados por el usuario en ficheros tipo documento o fichero tipo informe, además si algún fichero esta en dos columnas y el resto en una sola se va al traste la conversión. El programa no es tan estable como el resto suponemos que será porque ser versión Trial, pero una de cada tres conversiones el programa deja de funcionar lanzando un mensaje de error. En resumen no creo que aporte nada al mundo de la conversión pdf además cueste bastante dinero para lo que nos oferta.

Programa:	LD-Getter
Donde encontrarlo:	http://www.pdftoall.com/
Tipo:	Profesional
Licencia:	ShareWare (La licencia cuesta 71,85€ para un usuario, para empresa 1078.18€)
Presentación:	Se presenta una única versión para Windows, se trata de un plugin para el Acrobat de adobe por lo cual el coste es elevado. Para este estudio se usara la versión Trial de ambas programas.
Instalación:	Se presenta con un fichero zip donde podemos encontrar el fichero de instalación que es el típico asistente de Windows, se instala sin ningún problema y las opciones de configuración son inexistentes, todo esto siempre que tengas instalado el Acrobat sino solo se ve un mensaje de error.
Utilización:	Al tratarse de una instalación de asistente típica el programa lo encontraremos en el grupo de trabajo que nosotros hayamos escogido pero como se trata de un plugin estará incrustado en el propio Acrobat, dentro de esta carpeta sin embargo se encuentra la ayuda, la dirección web de contacto y el acuerdo de licencia que hemos escogido en este caso Trial. Para poder usar el plugin ejecutamos el Adobe Acrobat y lo encontramos bajo el SubMenú plugins.
Resultado obtenido:	No podemos realizar un análisis tan extenso como en los programas de versión libre puesto que al ser una versión trial, nos avisa de que algunos caracteres serán sustituidos por X, lo cual hace el texto bastante complicado de seguir, ya que a veces se tratan solo de letras a veces de palabras enteras. Extrae también los encabezados y los pies de pagina y en presencia de imágenes, deja hueco donde estas deberían ir aproximando el tamaño, replicando la información del título de las imágenes. La distribución del texto en columnas no le supone ningún problema, y lo presenta tal cual es el fichero de entrada. En cuanto a la página de referencias al intentar mantener el aspecto de dos columnas del texto de entrada ocasiona colisiones entre las líneas de ambas columnas haciendo un procesado complicado cuando se mezclan cuadros de texto, junto a texto plano.
Conclusiones:	Entorno profesional y de pago, es un plugin de un programa, no es un programa en sí, el gasto de recursos maquina nos resulta alto, ya que se trata del Acrobat que de por si consume bastantes recursos la conversión es rápida. El programa es estable pero solo me ha dejado ejecutarlo una vez la segunda me aparece como deshabilitado o me da un mensaje de error. En resumen convierte bastante bien si no se tiene en cuenta la cosa de los cambios por las X pero por mucho menor coste nos podemos encontrar con programas mucho más livianos y que hacen algo parecido o mucho mejor.

Programa:	LD-Getter Pro
Donde encontrarlo:	http://www.pdftoall.com/
Tipo:	Profesional
Licencia:	ShareWare (La licencia cuesta 158€ para un usuario)
Presentación:	Se presenta una única versión para Windows, se trata de un plugin para el Acrobat de adobe por lo cual viene a costar bastante dinero, para este estudio se usara la versión Trial de ambas programas. Se presenta como la versión Pro de LD-Getter
Instalación:	Se presenta con un fichero zip donde podemos encontrar el fichero de instalación que es el típico asistente de Windows, se instala sin ningún problema y las opciones de configuración son inexistentes, todo esto siempre que tengas instalado el Acrobat sino solo se ve un mensaje de error.
Utilización:	Al tratarse de una instalación de asistente típica el programa lo encontraremos en el grupo de trabajo que nosotros hayamos escogido pero como se trata de un plugin estará incrustado en el propio Acrobat, dentro de esta carpeta sin embargo se encuentra el cómo se usa y el acuerdo de licencia que hemos escogido en este caso Trial. Para poder usar el plugin ejecutamos el Adobe Acrobat y lo encontramos bajo el SubMenú plugins, escogemos el fichero pdf del cual vamos a partir y puedes darle información adicional de cómo es el fichero de entrada, le especificas la carpeta de salida.
Resultado obtenido:	No podemos realizar un análisis tan extenso como en los programas de versión libre puesto que al ser una versión trial, nos avisa de que solo convertirá las paginas impares. Extrae los encabezados y los pies de pagina y en presencia de imágenes, deja hueco donde estas deberían ir aproximando el tamaño, replicando la información del título de las imágenes. La distribución del texto en columnas no le supone ningún problema, y lo presenta tal cual es el fichero de entrada dejando un gran espaciado entre las columnas para su correcta lectura. En cuanto a la página de referencias estas son claras y fácilmente seguible además de procesables.
Conclusiones:	Entorno profesional y de pago, es un plugin de un programa, no es un programa en sí, el gasto de recursos maquina nos resulta alto, ya que se trata del Acrobat que de por si consume bastantes recursos la conversión es rápida. En resumen convierte bastante bien si no se tiene en cuenta la cosa de los cambios por las X pero por mucho menor coste nos podemos encontrar con programas mucho más livianos y que hacen algo parecido o mucho mejor. La única diferencia que se ha visto entre la versión profesional y la normal es que es un poco más rápida la versión profesional, pero no merece la pena pagar la diferencia de precio, entre uno y otro.

Programa:	PDF Plain Text Extractor
Donde encontrarlo:	http://www.retsinasoftware.com/
Tipo:	Profesional
Licencia:	ShareWare (La licencia cuesta 59,95€ para un usuario)
Presentación:	Se presenta una única versión para Windows, se trata de un fichero ejecutable que contiene la instalación del programa. La versión aquí analizada es: PDF Plain Text Extractor V4.0
Instalación:	Se presenta con un fichero zip donde podemos encontrar el fichero de instalación que es el típico asistente de Windows, se instala sin ningún problema y las opciones de configuración son inexistentes.
Utilización:	Al tratarse de una instalación de asistente típica el programa lo encontraremos en el grupo de trabajo que nosotros hayamos escogido se encuentra la ayuda y el acuerdo de licencia que hemos escogido en este caso Trial. El uso es un poco diferente al resto de los anteriores, con el propio explorador de ficheros que tenemos a nuestra izquierda buscaremos el fichero pdf a convertir y luego lo arrastraremos hasta la parte derecha de la ventana, si clickeamos sobre el nos mostrara la meta-información del fichero. Podemos seleccionar algunas opciones como rangos de páginas o la inserción de marcadores de página.
Resultado obtenido:	No podemos realizar un análisis tan extenso como en los programas de versión libre puesto que al ser una versión trial, nos avisa de que solo convertirán las cinco primeras páginas Lo primero que hace es insertar la meta información del fichero en la conversión, opción por defecto que hay que desmarcar. Extrae los encabezados y los pies de pagina y en presencia de imágenes, deja hueco donde estas deberían ir aproximando el tamaño, replicando la información del título de las imágenes. La distribución del texto en el fichero de salida intenta adecuarse al fichero de entrada pero enseguida la pierde incluso llega a mezclar cosas de una página dentro de otra que no le debería corresponder. En cuanto a la página de referencias no lo podemos saber puesto que no nos permite llegar a ella con su conversión.
Conclusiones:	Entorno agradable y atrayente con botones intuitivos aunque la forma de añadir los ficheros al conversor no es muy intuitiva ya que hay que arrastrar los ficheros a su ventana en vez de cargarlos mediante asistentes, puede hacer conversión en serie de varios ficheros y lo hace bastante bien y rápido, sin un uso excesivo de recursos. En cuanto a la calidad de la conversión no es demasiado buena ya que mezcla cosas de una página con otras, no es muy recomendable este comportamiento

2.1.3 CONCLUSIONES

En un primer acercamiento a las herramientas de extracción que hemos utilizado, vemos que las versiones comerciales de los programas no aportan ninguna característica extra a las versiones gratuitas.

Los resultados obtenidos varían bastante dependiendo del programa utilizado. En nuestro caso destacaríamos el xpdf por su versatilidad y calidad de los resultados, que se ajustan más a los resultados esperados. Sin embargo al carecer de una interfaz gráfica podría no gustar a determinados usuarios que no se sienten cómodos con el uso de la consola para realizar las operaciones.

El formato de las publicaciones científicas suele dividir el texto en 2 columnas, lo que dificulta el trabajo a las herramientas. Algunas de ellas mantienen la estructura de las columnas visualmente en el archivo de texto y esto una desventaja ya que la conversión se hace principalmente para el procesado del texto. Otros sin embargo mantienen la linealidad del texto eliminando formatos de presentación.

Otra característica de algunos programas es que dividen los resultados en diferentes archivos, es decir crean un archivo diferente por cada página del PDF original. Esto podría ser útil para ciertas aplicaciones pero en general no para nuestro objetivo, por lo que trataremos de evitar esta característica.

Con este primer análisis superficial podemos ir viendo cuales serían algunos de los mejores programas, siendo el XPDF el favorito a pesar de no contar con una interfaz gráfica pero ofreciendo unos resultados muy buenos. Por otra parte los programas con licencias comerciales en general muestran unos resultados bastante decepcionantes y por debajo del nivel de programas gratuitos.

2.2 SEGUNDO ESTUDIO

2.2.1 CONSIDERACIONES INÍCIALES

En esta segunda prueba vamos a realizar un estudio más detallado de cada uno de los programas, señalando las funcionalidades que pose cada uno para hacer una comparativa. La comparativa servirá para ver de forma sencilla cuales son los programas que se adaptan a nuestras necesidades.

Para hacerlo de forma visual vamos a utilizar una tabla en la que se muestran por un lado los programas que vamos a comparar y las características de cada uno de ellos por otra parte. Las características elegidas son:

- **Licencia**
  Se muestra el tipo de licencia del programa, si es freeware, shareware u open source. Lo ideal sería una licencia opensource para poder adaptar el código a nuestras necesidades y optimizar los resultados.

- **Precio**

- **Programa / plugin Acrobat**
  Algunos de los programas necesitan tener instalado el Adobe Acrobat, esto obliga a pagar la licencia del mismo ya que es un programa comercial.

- **Automatización Mediante Scripts**
  Hemos considerado que la automatización mediante scripts es positivo ya que la posiblidad de creación de estos scripts puede hacer que el trabajo se vea reducido de una manera considerable ya que solo consistiría en la programación de el script y la ejecución de este mismo. El no tener esta funcionalidad hace que el usuario deba permanecer de manera exclusiva con este trabajo.

- **Mantiene el *layout***
  Mantiene el formato de presentación del archivo original en relación a las columnas. Lo mejor sería poder seleccionarlo pero en nuestro caso es suficiente que no lo mantenga y muestre el resultado en orden. Ciertos programas mantienen el formato de columna lo que hace casi ilegible el texto para una máquina.

- **Seleccionar páginas** Puede ser una función muy útil en determinados casos seleccionar el principio y fin del archivo que queremos convertir y ahorrar trabajo posterior.

- **Seleccionar codificación de caracteres** Aunque la mayoría de las publicaciones científicas son in inglés, puede resultar necesario hacer la conversión a tipos de caracteres especiales para documentos en otros idiomas con caracteres especiales.

- **Resultados en un solo archivo** Los documentos que vamos a tratar preferiblemente estarán en un mismo archivo, ya que las diferentes páginas del PDF original no atienden a cuestiones de contenido si no una simple cuestión espacial y de presentación.

- **Fin de línea** Aunque esta característica no es demasiado relevante, existen ciertos programas que tienen problemas para mostrar archivos con el fin de línea en un formato distinto al del sistema operativo que se está manejando.

2.2.2 Resultados Obtenidos

	Licencia	Precio	Programa / plugin Acrobat	Automatizacion mediante Scripts	Mantiene layout	Seleccionar paginas	Selección codificación de caracteres	Resultados en un solo archivo	Fin de línea
XPDF	GNU	0	programa	Si	Opcional	Si	Si	Si	unix | dos | mac
Cool PDF Reader	Freeware	0	programa	No	No	No	No	Un archivo por pagina	dos
Easy PDF to Text converter	Freeware	0	programa	No	No	No	No	Un archivo por pagina	dos
A-PDF Text Extractor	Freeware	0	programa	No (Version Freeware)	No	Si	No	Si	dos
PDF2TEXT	Shareware	29'95 $	plugin	No	No	No	No	Si	dos
LD GETTER	ShareWare	71,85 €	plugin	No	Si	No	No	Si	dos
LD GETTER PRO	ShareWare	158 €	plugin	No	Si	No	No	Si	dos
PDF Plain Text Extractor	ShareWare	59,95 €	programa	No	Si	No	No	Si	dos

2.2.3 CONCLUSIONES

En la tabla podemos ver como hay un programa que destaca con diferencia por encima del resto, el XPDF que a pesar de no tener una interfaz gráfica para facilitar la tarea de extracción de los textos, presenta la posibilidad de la realización de Scripts, es el más completo con muchísimas opciones de configuración para una mejor adaptación de los resultados a las necesidades. Además es un programa de código abierto y con licencia GNU, lo que permite modificar el código a tus necesidades libremente.

Otro programa interesante sería el A-PDF Text Extractor. Cuenta con una sencilla y muy usable interfaz gráfica para seleccionar los archivos para la extracción y poder configurar algunas opciones. Las opciones que permite configurar están relacionadas con la selección de las páginas, siendo el mejor en este aspecto aunque se echan en falta muchas otras opciones que podemos encontrar con el XPDF.

El resto de programas por lo general muestran unos resultados muy pobres y con poca calidad. La mayoría se limitan a interfaces gráficas donde poder seleccionar el archivo y convertirlo a texto pero sin dar ninguna opción de configuración. Además la usabilidad para una tarea tan sencilla como esa deja bastante que desear e incluso es complicada como por ejemplo con el Cool PDF READER.

Para el siguiente análisis vamos a utilizar sólo el XPDF y el A-PDF Text Extractor, descartando todos los demás ya que no aportan nada que no sea posible hacer con estos 2 programas escogidos.

2.3 TERCER ESTUDIO

2.3.1CONSIDERACIONES INÍCIALES

ASPECTOS TENIDOS EN CUENTA DURANTE LA TERCERA EVALUACIÓN:

En esta tercera y última prueba ya solo se tendrán en cuentan dos de los ocho programas iníciales. Estos dos programas son los que bajo nuestro criterio son los que mejor se han comportado a la hora de la extracción del texto de los ficheros PDF.

En este tercer estudio nos fijaremos de una manera intensiva en cuestiones relacionadas al rendimiento y la carga de trabajo que la ejecución de estos dos programas provoca en la maquina en la que corren, esta máquina será detallada en la parte del Banco de Trabajo como maquina referencia.

Para mostrar dicho impacto se capturaran los momentos en los que la maquina está trabajando de modo no exclusivo en esta tarea, es decir se tendrán abiertas diversas aplicaciones para simular un entorno de trabajo lo más real posible, se mantendrán abiertas aplicaciones tipo procesador de texto (WORD), aplicaciones tipo navegador web (Mozilla Firefox 3.0), aplicaciones tipo Messenger y es probable que también se incluya la presencia de reproductores multimedia.

Estas capturas ser realizaran mientras se esté trabajando en los 7 ficheros pdf que se marcaron en el banco de trabajo. La selección de estos ficheros se corresponde al siguiente criterio: dos de ellos son los presentes durante todo el trabajo, mientras que los otros cinco serán escogidos siguiendo diversos criterios en cada uno de ellos, tales como peso del fichero, el que contenga o no contenga imágenes, extensión del texto presente en el, formateado en una o dos columnas este texto.

También vamos a evaluar la carga de trabajo que supone para el usuario la conversión de una batería de ficheros pdf, estos ficheros serán más detallados en la sección banco de trabajo, fijándonos en una variable que es la productividad, que la obtendremos de la siguiente manera:

$$\frac{\text{Tiempo empleado}}{\text{Trabajos Realizados}}$$

El trabajo a realizar consiste en realizar un conjunto de conversiones con ambos programas y medir los tiempos que se tarda en realizar dichas conversiones.

El tiempo empleado se mide en segundos.

Esta medida obtenida en trabajo por segundo estimará el tiempo medio en la ejecución de un trabajo, que cuanto más baja sea indicará que mejor es la aplicación puesto que se supone un empleado ideal al que no le afecten variables de entorno, o fatiga. Aunque parezca una medida irrelevante, en programas con entorno gráfico no susceptible a Scripts o automatización la atención del trabajador a la tarea es algo que influye de manera muy fuerte en el trabajo a reliazar.

Para el XPDF se usarán dos tipos de estrategia de conversión la primera de ellas, consistirá en realizarlo de una forma manual a través de la línea de comandos y la segunda manera consistirá en realizarlo mediante la técnica de Drag and Drop (Se explica mas adelante en que consiste); para el APDF lo único que podemos comprobar es la conversión mediante la interfaz grafica, puesto que la versión consola es de pago y no la podemos testear.

La carga de trabajo de la que se va a disponer es de 27 ficheros pdf.

BANCO DE TRABAJO.

MAQUINA REFERENCIA:

PROCESADOR:

AMD ATHLON 64 PROCESSOR 3000+

CACHE:

2 Niveles de cache ONCHIP
Primer nivel de 64KB
Segundo nivel de 512KB

2 Niveles de cache ONBOARD
Primer nivel de 1 MB
Segundo nivel de 2 MB

MEMORIA RAM:

1.5GB Montado en tres bancos 1GB, /256MB/256MB, 200MHz

DISCO DURO:

160GB IDE 133

SISTEMA OPERATIVO:

Microsoft Windows XP (Service Pack 3)

PROGRAMAS QUE SERÁN ANALIZADOS:

- XPDF.
- APDF Text Extractor

FICHEROS DE PRUEBA:

Para un manejo más sencillo todos los ficheros han sido renombrados a ficheroxx.pdf donde xx es un numero correlativo entre 1 y 27, no obstante en la información posterior se mostrara un enlace web a cada una de las paginas correspondientes al fichero, y una marca si dicho fichero ha sido seleccionado para la evaluación de rendimiento.

Fichero 01	Seleccionado
http://www.biomedcentral.com/1471-213X/5/14	SI

Fichero 02	Seleccionado
http://dev.biologists.org/cgi/reprint/133/10/1901	SI

Fichero 03	Seleccionado
http://www.molecularbrain.com/content/1/1/4	SI

Fichero 04	Seleccionado
http://www.biomedcentral.com/1471-2105/9/292	NO

Fichero 05	Seleccionado
http://www.biomedcentral.com/1471-2105/9/286	NO

Fichero 06	Seleccionado
http://www.biomedcentral.com/1471-2105/9/284	NO

Fichero 07	Seleccionado
http://www.biomedcentral.com/1471-2105/9/266	NO

Fichero 08	Seleccionado
http://www.biomedcentral.com/1471-2105/9/206	NO

Fichero 09	Seleccionado
http://genomebiology.com/2008/9/8/R132	NO

Fichero 10	Seleccionado
http://arthritis-research.com/content/10/4/R98	NO

Fichero 11	Seleccionado
http://www.biomedcentral.com/1471-2164/9/389/abstract	SI

Fichero 12	Seleccionado
http://genomebiology.com/2008/9/8/R128	NO

Fichero 13	Seleccionado
http://www.biomedcentral.com/1752-0509/2/73/abstract	SI

Fichero 14	Seleccionado
http://www.cardiab.com/content/7/1/24	NO

Fichero 15	Seleccionado
http://www.biomedcentral.com/1471-2164/9/379/abstract	NO

Fichero 16	Seleccionado
http://genomebiology.com/2008/9/8/R123	SI

Fichero 17	Seleccionado
http://www.biomedcentral.com/1471-2148/8/226/abstract	NO

Fichero 18	Seleccionado
http://www.biomedcentral.com/1471-2164/9/361/abstract	NO

Fichero 19	Seleccionado
http://dev.biologists.org/cgi/reprint/133/10/1871	NO

Fichero 20	Seleccionado
http://dev.biologists.org/cgi/reprint/133/10/1891	SI

Fichero 21	Seleccionado
http://dev.biologists.org/cgi/reprint/133/10/1911	NO

Fichero 22	Seleccionado
http://dev.biologists.org/cgi/reprint/133/10/1933	NO

Fichero 23	Seleccionado
http://dev.biologists.org/cgi/reprint/133/10/1955	NO

Fichero 24	Seleccionado
http://dev.biologists.org/cgi/reprint/133/10/1979	NO

Fichero 25	Seleccionado
http://dev.biologists.org/cgi/reprint/133/10/2001	NO

Fichero 26	Seleccionado
http://dev.biologists.org/cgi/reprint/132/1/35	NO

Fichero 27	Seleccionado
http://dev.biologists.org/cgi/reprint/132/1/75	NO

2.3.2 RESULTADOS OBTENIDOS.

La primera parte de los resultados se corresponde al estudio de consumos tanto de CPU como de memoria RAM por los dos programas escogidos.

Ambos programas se ejecutaran en igualdad de condiciones es decir sobre la misma máquina con una misma carga de trabajo y manteniendo el Layout del texto de entrada.

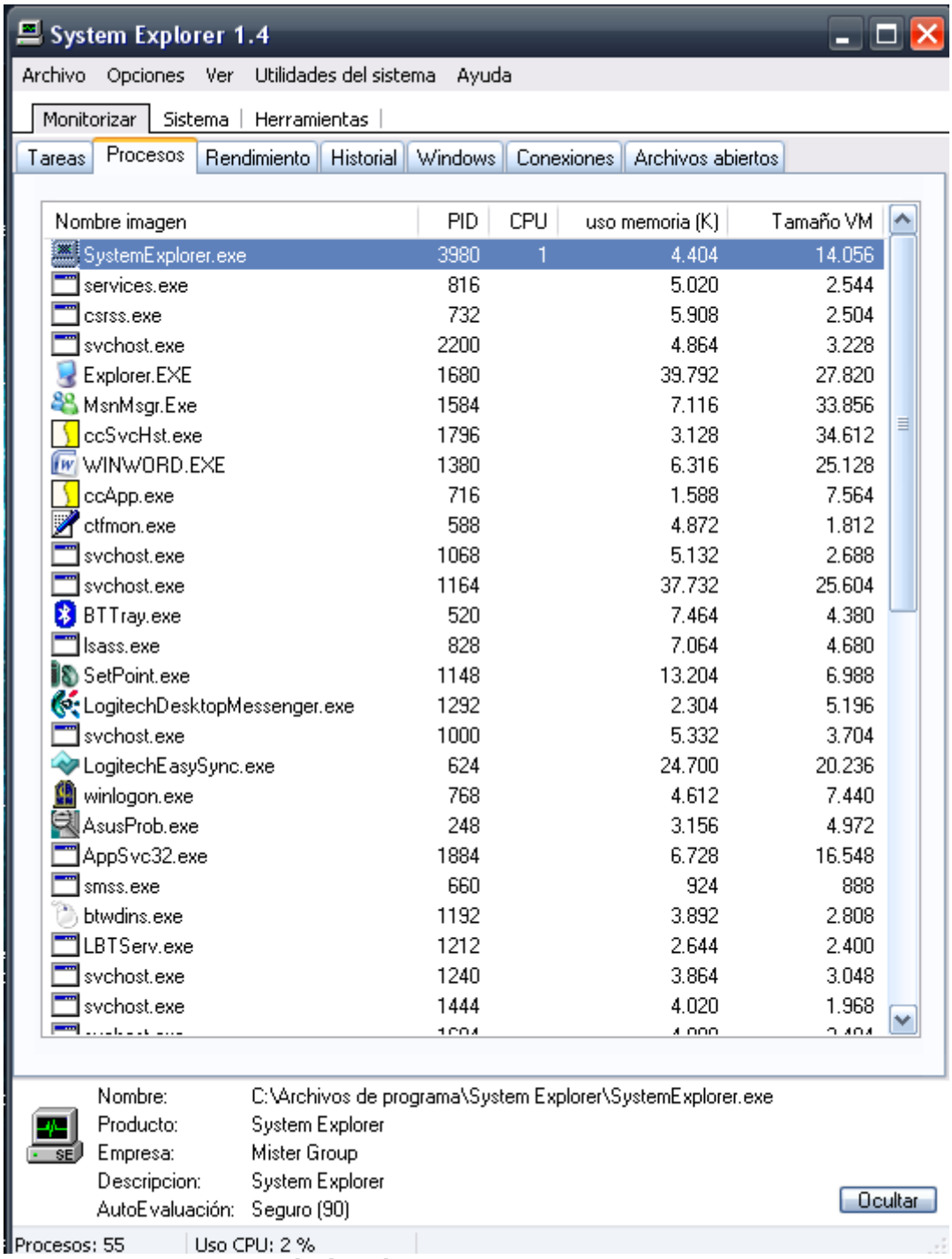


Figura 1. Maquina en estado de trabajo estacionario.

La segunda parte será el estudio de productividad de los programas.

XPDF:

1ª Parte:

Para poder analizar el rendimiento con este programa, hay que tener en cuenta que no se crea un proceso único y hay que comprobar varios. Para identificar los procesos ejecutados durante la conversión, se comprueban aquellos que hacen uso de la CPU, aquellos que se ven implicados muestran un consumo elevado en el procesador.

Las conversiones se realizaran mediante un proceso llamado Drag and Drop consistente en volcar el fichero que deseamos convertir o extraer el texto, encima del icono de la aplicación (no exactamente el icono sino un acceso directo para poder concatenarle al paso de parámetros, el flag de conservación de layout).

Al realizar una conversión se ha observado que los procesos que entran en juego son, el propio XPDF cuando el fichero a convertir es de gran volumen, (en ficheros de bajo volumen la presencia del XPDF es prácticamente minima puesto que la vigencia de la instancia es muy baja) el propio explorador de ficheros de Windows creando el que será el fichero de salida, y el servidor de aplicaciones locales.

Para cada una de los ficheros del banco de pruebas se mostrará la siguiente información que será común para ambos programas: el nombre del fichero, el peso en MBytes o en KiloBytes de dicho fichero, un breve resumen del contenido del fichero, y una breve valoración de los resultados obtenidos. (Véase tabla 1)

Este sistema de evaluación será idéntico para el otro programa implicado.

Nombre:	Nombre fichero	Tamaño:	Tamaño KB o MB
Comentario acerca del fichero			
Captura obtenida			
Comentario de los resultados			

Tabla 1.

Y al final de cada conjunto de capturas se presentara una tabla resumen con la información obtenida de dichas capturas.

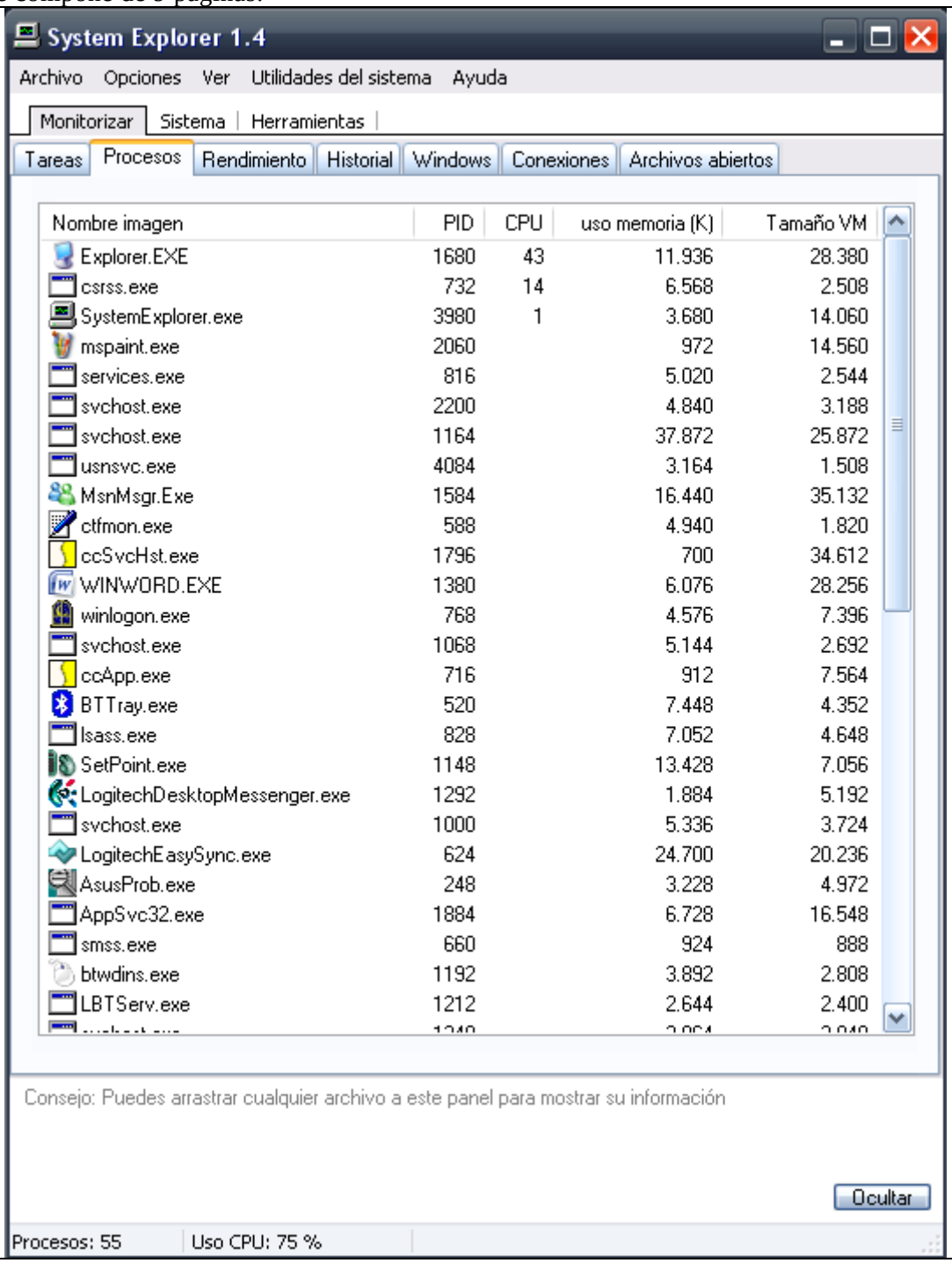
Resultados:	
Numero de Fichero	Consumo de CPU acumulado
	Consumo de Memoria acumulado

Tabla 2.

Cada fila doble de la Tabla 2 se corresponderá a un fichero del banco de pruebas

Nombre:	Fichero01.pdf	Tamaño:	2.06 MB

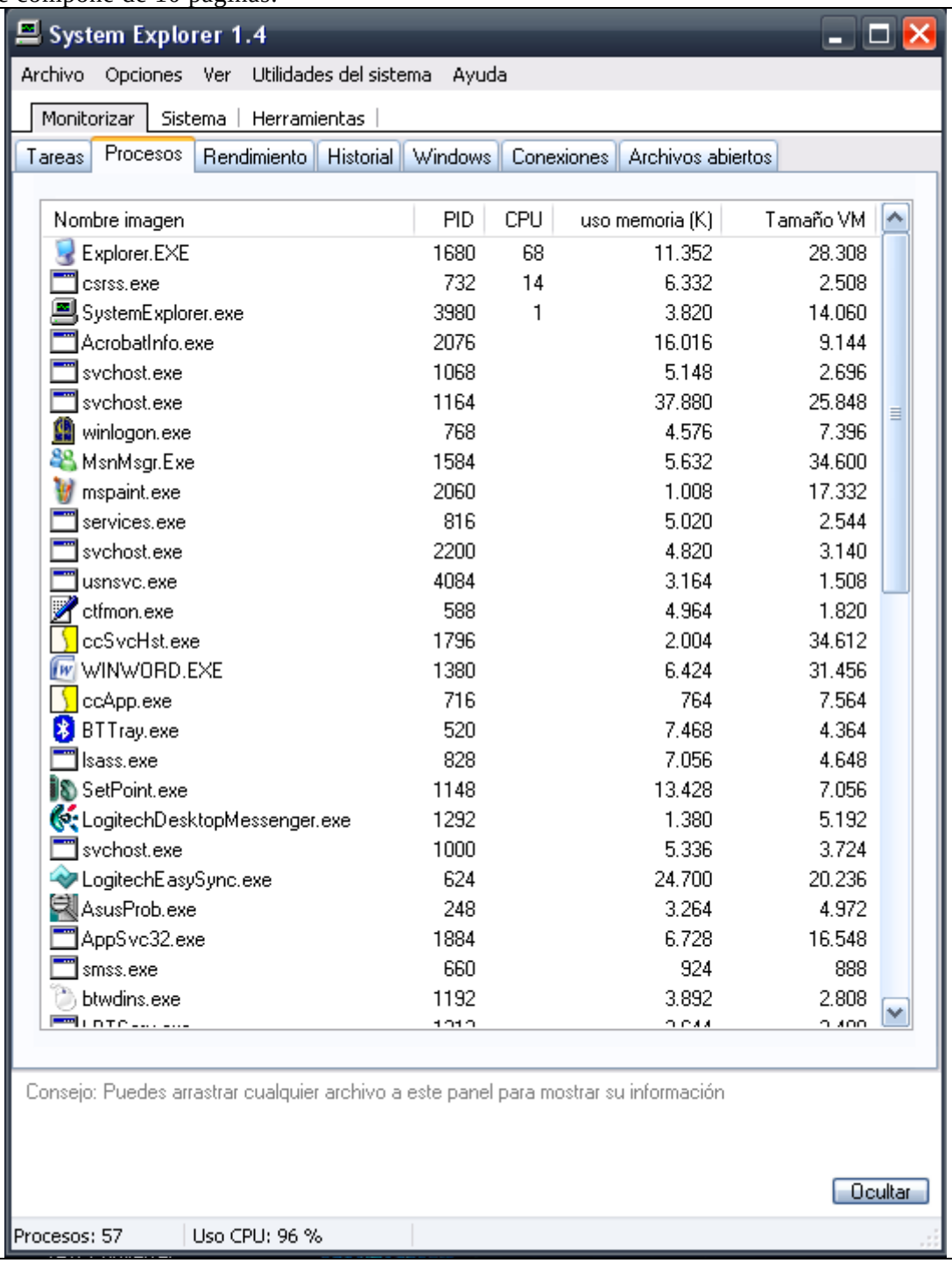
Este fue el fichero de referencia para la primera prueba.
Se puede considerar un fichero tipo, presenta el Abstract a una columna y el resto del fichero a dos columnas.
Incluye imágenes de un tamaño pequeño.
Se compone de 9 páginas.



El único consumo que se aprecia es el de los programas auxiliares que son el explorador de Windows y el servidor de aplicaciones de Windows, el propio extractor no aparece si quiera puesto que el fichero es sencillo y suficientemente ligero, como para que no le presente demasiada complicación.

Nombre:	Fichero02.pdf	Tamaño:	1.11 MB

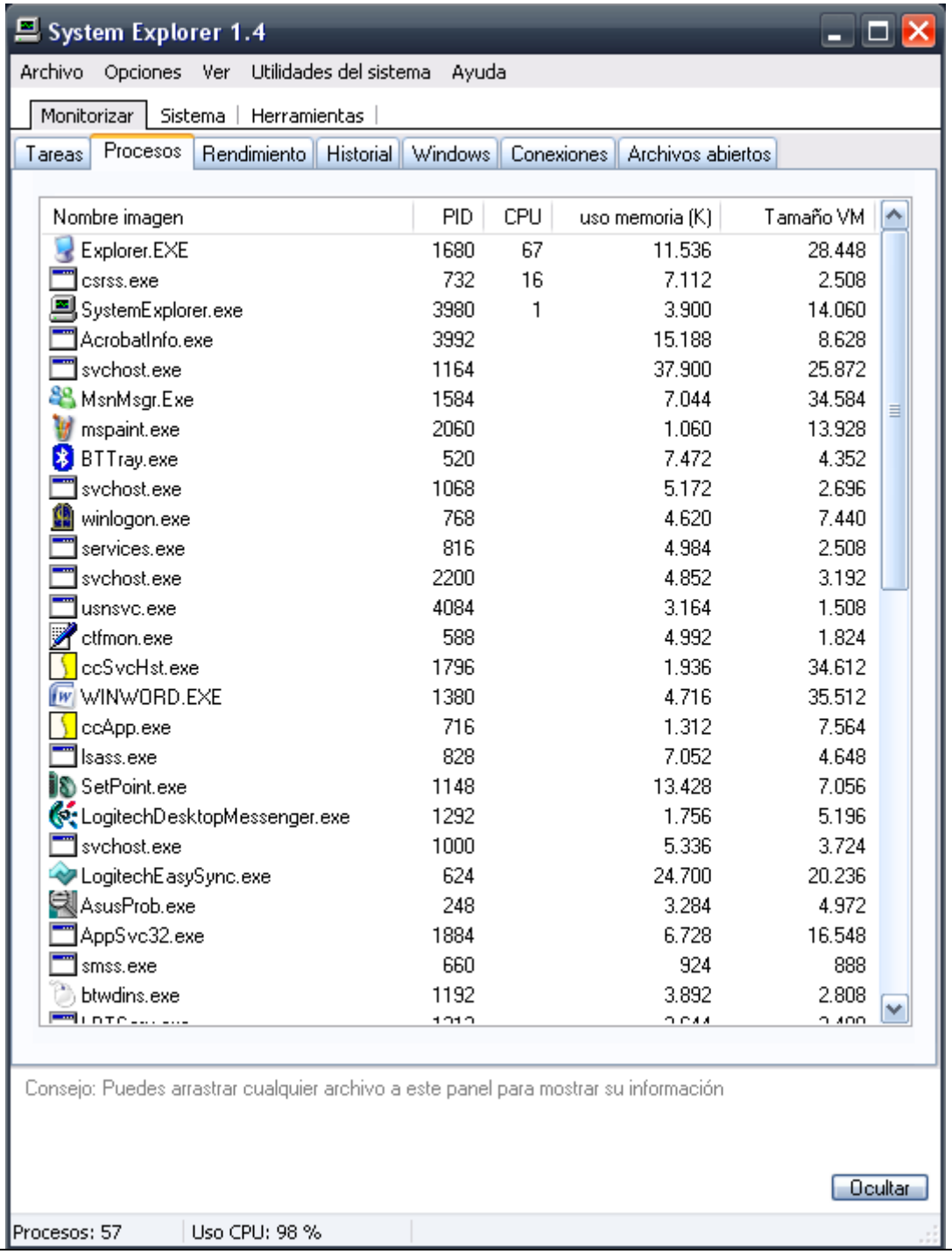
Se puede considerar un fichero tipo, presenta Abstract pero no identificado como tal a una columna y el resto del fichero a dos columnas.
Incluye imágenes de tamaños variables.
Se compone de 10 páginas.



El único consumo que se aprecia es el de los programas auxiliares pero de una manera más notable, el fichero de salida es sensiblemente más grande en tamaño, el propio extractor no aparece si quiera puesto que el fichero es sencillo y suficientemente ligero, como para que no le presente demasiada complicación.

Nombre:	Fichero03.pdf	Tamaño:	178 KB

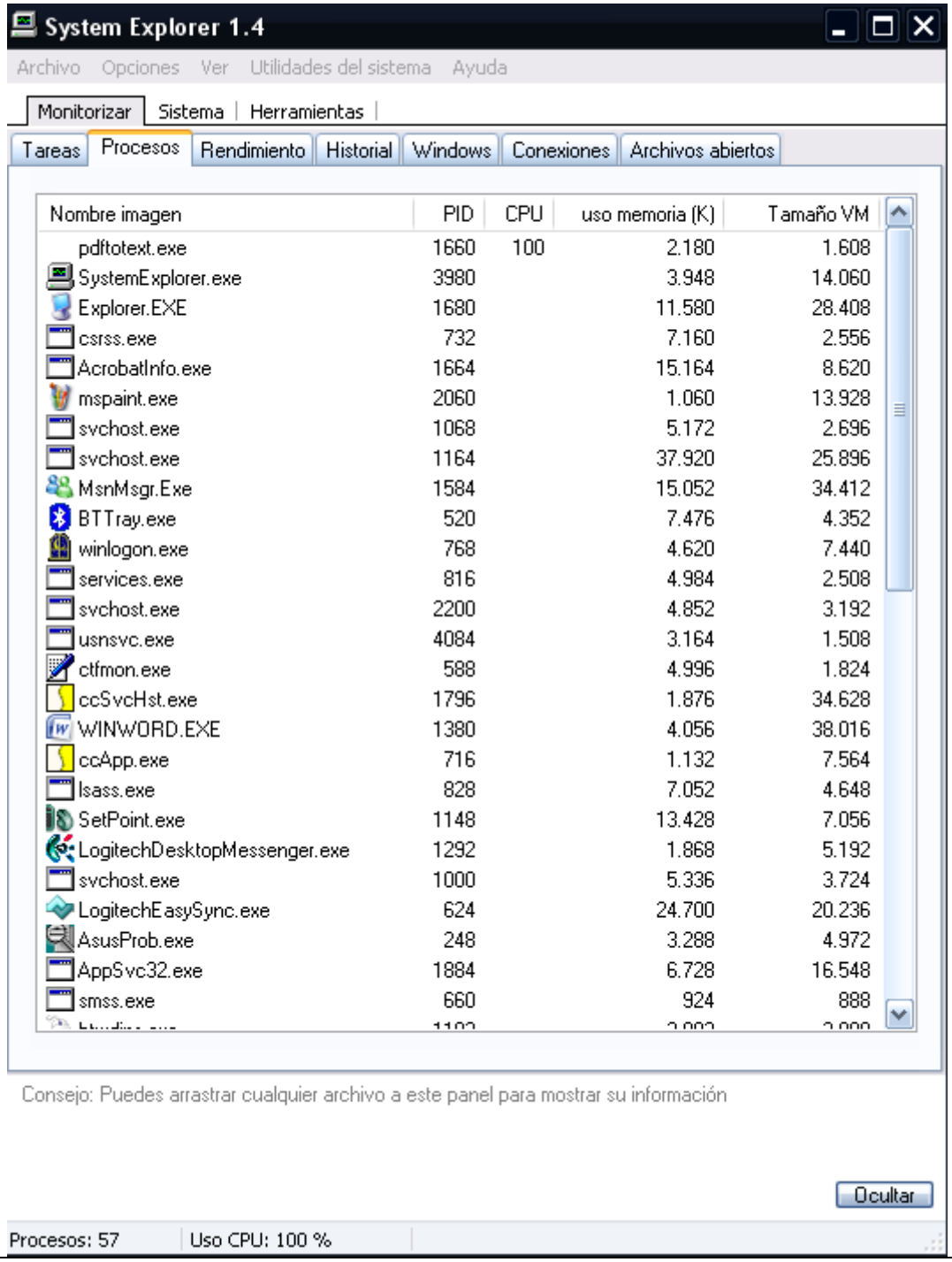
Se trata de un artículo aun no completo no acorde al formato estándar.
Contiene todas las imágenes al final del texto.
Se compone de 19 páginas.



El único consumo que se aprecia es el de los programas auxiliares pero de una manera más notable él, el propio extractor no aparece si quiera puesto que aun no estando acorde a un fichero estándar, el extractor no distingue este tipo de cuestiones.

Nombre:	Fichero11.pdf	Tamaño:	4.92 MB

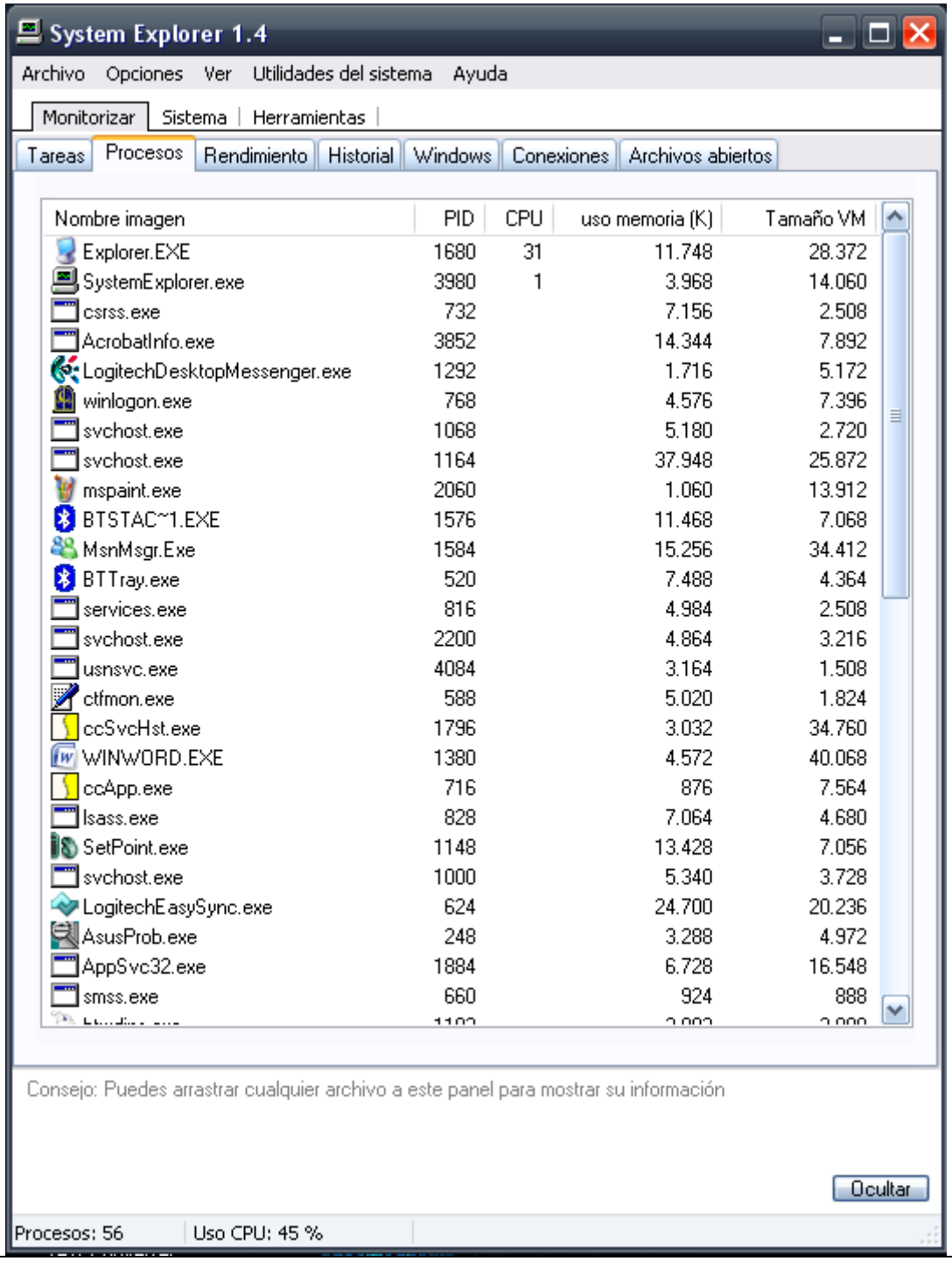
Se trata de un artículo aun no completo no acorde al formato estándar.
Contiene todas las imágenes al final del texto, se tratan de unas imágenes muy ricas en datos puesto que son formas de onda y son altamente detalladas.
Se compone de 36 páginas.



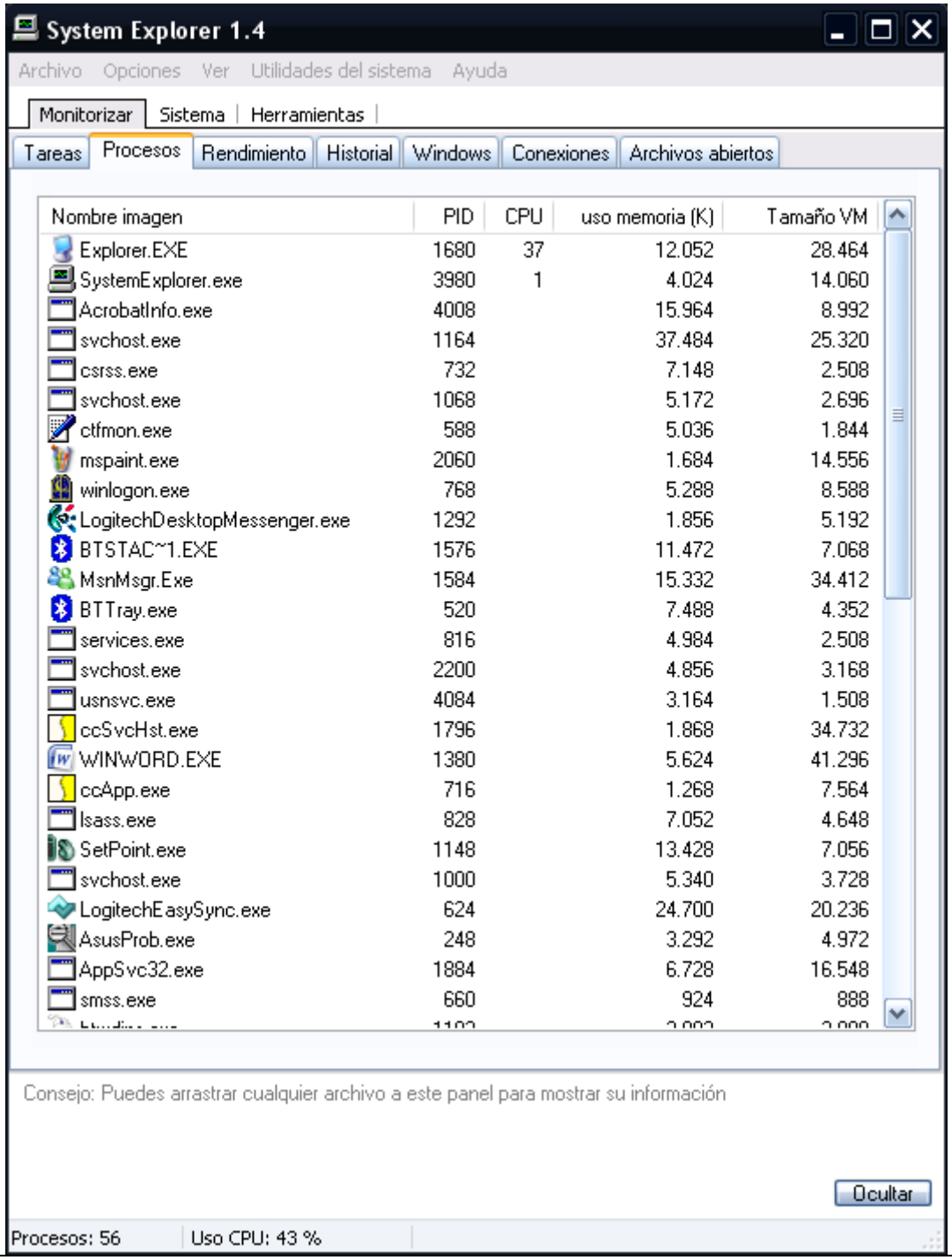
El primer fichero que le presenta un reto al programa, como se puede comprobar llega a cargar la CPU hasta un 100% y requiere de uso de memoria RAM unos 2 Megas, durante un periodo de unos 2 segundos, se puede concluir que cuando el fichero presenta imágenes a mucha resolución el proceso se ve sensiblemente afectado.

Nombre:	Fichero13.pdf	Tamaño:	230 KB

Se trata de un artículo aun no completo no acorde al formato estándar.
Contiene todas las imágenes al final del texto, se tratan de unas imágenes muy simples, son unos esquemáticos también presenta tablas de datos.
Se compone de 17 páginas.



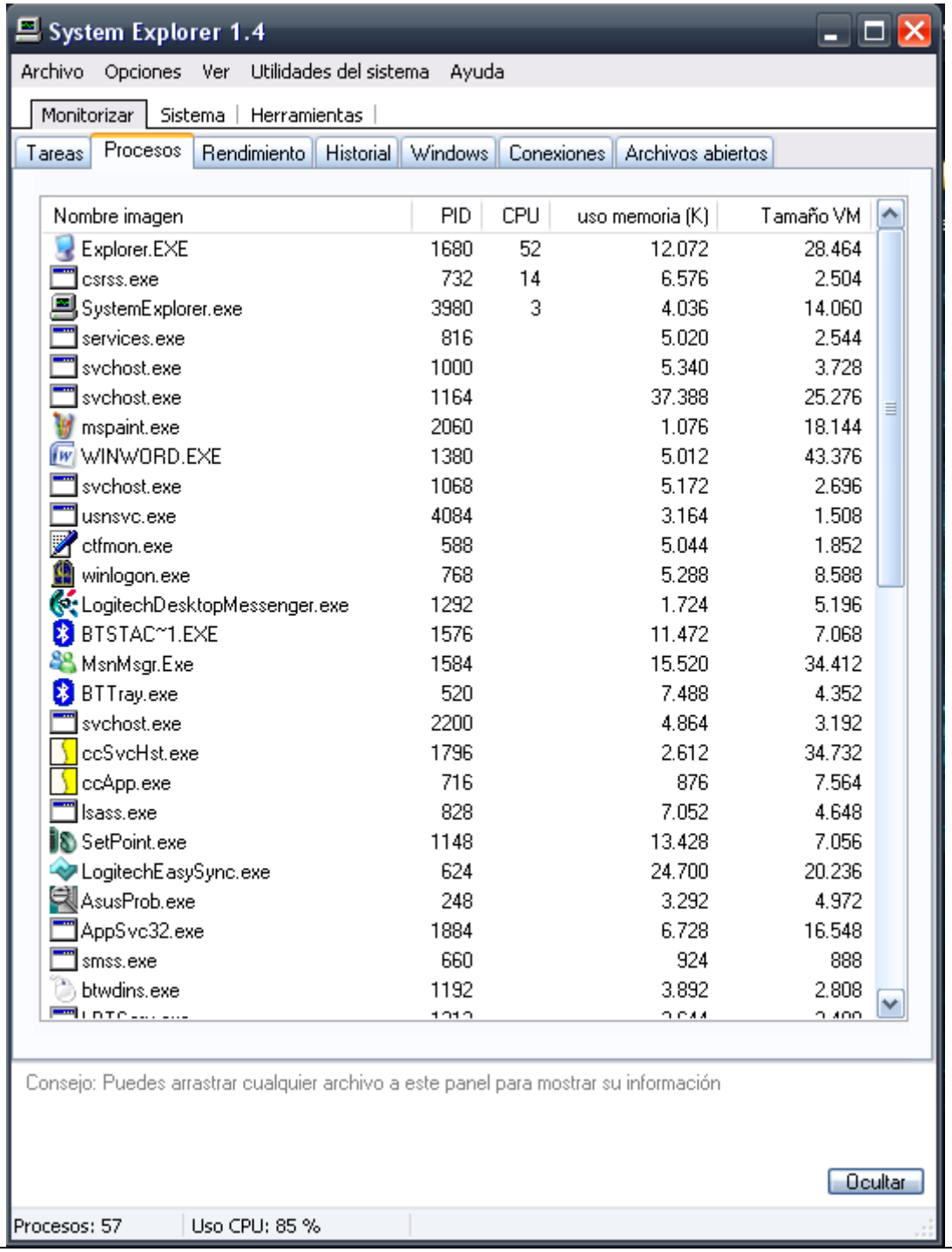
Fichero simple y sin ninguna complicación, presenta un consumo de recursos moderado bajo.			
Nombre:	Fichero16.pdf	Tamaño:	793 KB
Se puede considerar un fichero tipo, presenta el Abstract a una columna y el resto del fichero a dos columnas. Incluye imágenes de diverso tamaño una de ellas especialmente rica. Se compone de 9 páginas.			



Se aprecia un consumo bajo de los procesos auxiliares, pero esta conversión resulto más compleja de lo que aquí se muestra, ya que se creó una instancia completa del XPDF, con una vigencia de un segundo aproximadamente. Se puede empezar a observar que cuando las imágenes son ricas el programa consume muchos más recursos.

Nombre:	Fichero20.pdf	Tamaño:	5.74 MB

Se puede considerar un fichero tipo, presenta el Abstract a una columna y el resto del fichero a dos columnas.
Incluye imágenes de diverso tamaño son medianamente ricas en cuanto a detalles, se prevé que el programa presente carga de trabajo.
Se compone de 10 páginas.



Se vuelve a ver un consumo alto de procesos auxiliares, y también existió una instancia del XPDF con una vigencia lo suficientemente larga como para ser tenida en cuenta.

Tabla Resumen:	
Fichero01.pdf	57%
	18,522 MB
Fichero02.pdf	82%
	17,684MB
Fichero03.pdf	83%
	18,684 MB
Fichero11.pdf	100%
	2,180 MB
Fichero13.pdf	34%
	11.748 MB
Fichero16.pdf	38%
	16,076MB
Fichero20.pdf	66%
	18,648 MB

A la vista de estos resultados obtenidos se puede concluir las siguientes afirmaciones:

La extensión en páginas en el fichero no es un motivo de consumo suficientemente significativo, la única afirmación que se puede hacer es que a más paginas, más tiempo de conversión, lo cual resulta obvio.

El peso de los ficheros tampoco es determinante pero sí que puede avisar de lo que pude contener, es decir si el fichero contiene pocas páginas pero es muy pesado eso implica que tiene imágenes muy ricas por lo cual si que se ve forzado a mas carga de trabajo, y por otro lado si el fichero tiene muchas páginas y poco peso, esto implicara que tiene pocas imágenes ricas por lo cual el proceso no se verá muy alterado.

Como se puede deducir del párrafo anterior el consumo de recursos a nivel CPU se dispara en cuanto el fichero contiene imágenes con una gran resolución o una gran cantidad de datos, no es porque el programa intente extraer datos de ahí, sino que tiene que procesar gran cantidad de información que no es útil para la conversión final.

En cuanto al uso de memoria RAM nunca se ha visto descontrolado.

Concluyendo, el programa presenta un comportamiento bueno y no es gran consumidor de recursos.

2ª Parte:

Estudio de productividad.

Resultados de la primera estrategia, Drag and Drop:

Tiempo: 2 Minutos y14 Segundos lo que es un total de 124 segundos.
El número de tareas es de 27.

Total de productividad= 124 segundos /27 Tareas = 4.59 segundos por tarea.

Resultados de la segunda estrategia, conversión manual:

Tiempo: 2 Minutos 42 Segundos lo que es un total de 162 segundos.
El número de tareas es de 27.

Total de productividad= 162 segundos /27 Tareas = 6 segundos por tarea.

Viendo los resultados aquí obtenidos se observa que la conversión Drag and Drop nos da unos resultados en promedio de 1 segundo más rápida cada conversión, extrapolando estos términos se ve que cada 60 ficheros convertidos de la manera Drag and Drop nos ahorramos 1 minuto frente a la manera en serie.

Suponiendo que tuviésemos un trabajador ideal sin fatiga trabajando 8 horas en exclusivo a esta tarea, lo que hace un total de 28.800, segundos trabajando de la primera manera obtendríamos que convertiría 6.274 ficheros en esas 8 horas, mientras que usando la manera manual obtendríamos 4.800 ficheros, lo cual nos muestra una diferencia de unos 1.400 ficheros. Lo que en términos de productividad son unos datos bastante a tener en cuenta.

A-PDF Text Extractor:

1ª Parte:

Este programa al tratarse de una aplicación para Windows es sencillo de controlar ya que al ejecutarse lanza un proceso completamente etiquetado e identificable en la lista de procesos del sistema.

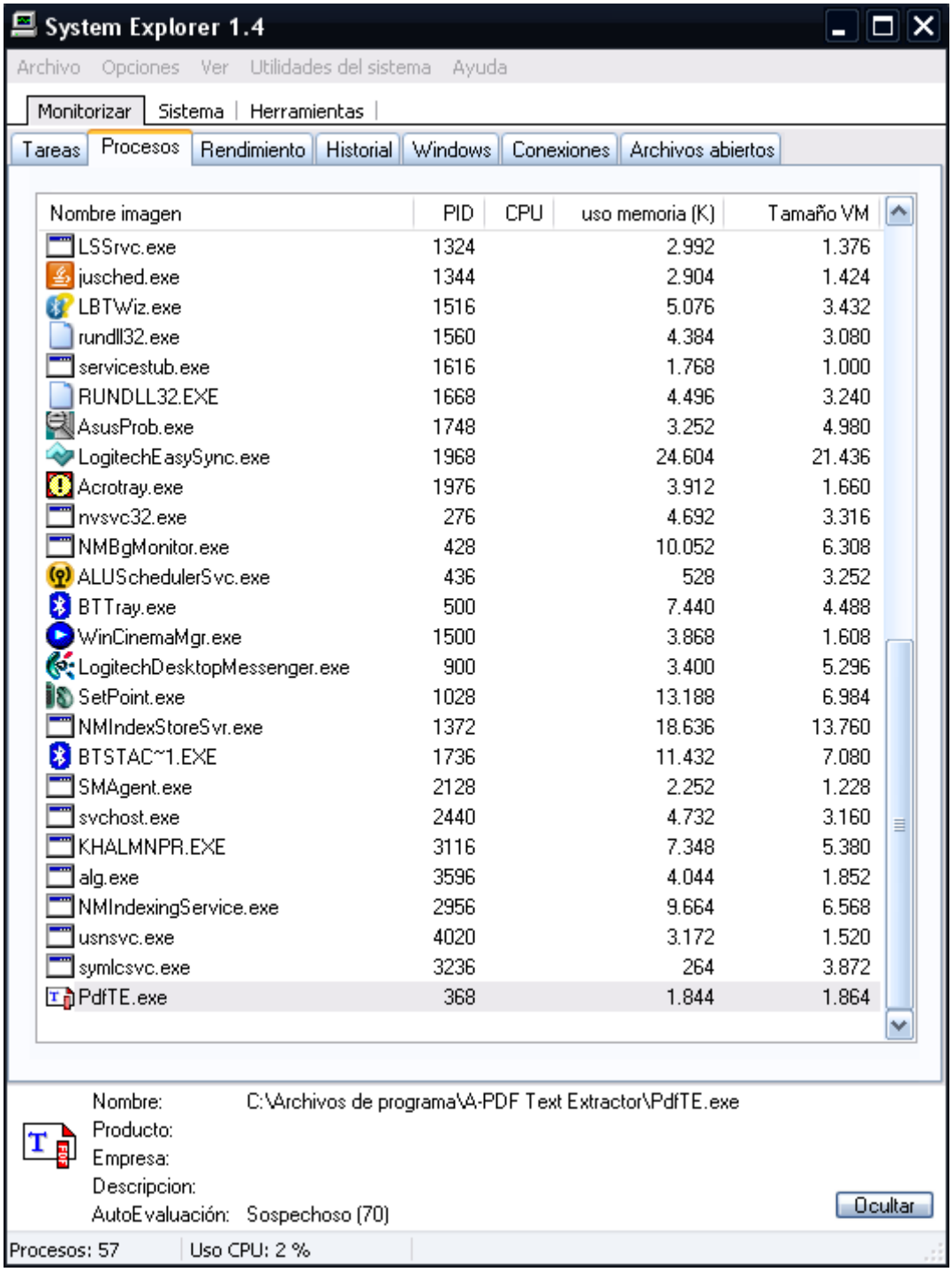


Figura 2. Nueva definición de estado estacionario.

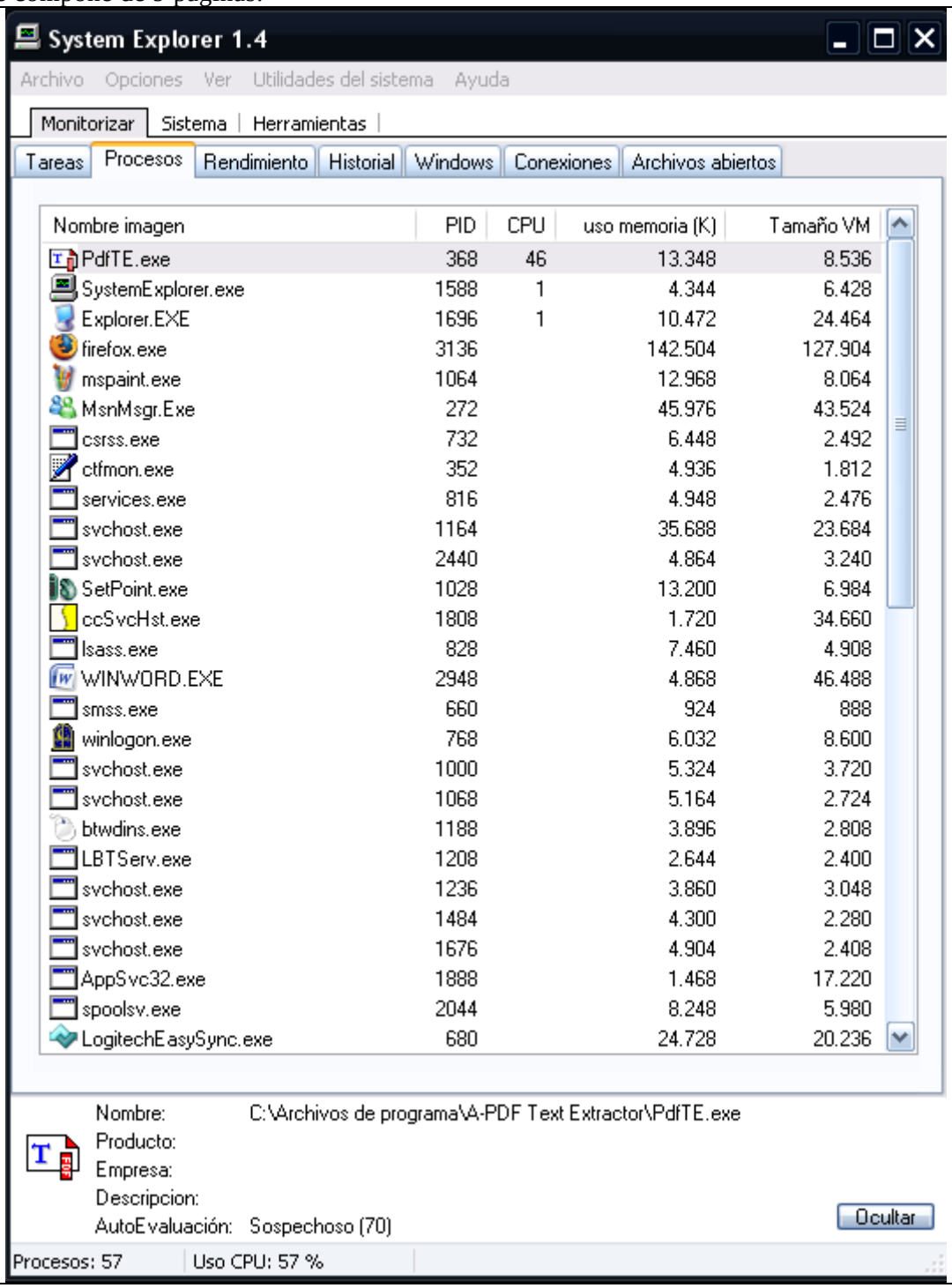
Al permanecer en ejecución el estado estacionario como se definió para el caso del XPDF no es suficientemente bueno, ya que en ese estado no se tiene constancia del estado residente de este programa. Por lo cual se muestra una captura del estado estacionario con el A-PDF Text Extractor en ejecución, e inmediatamente después ya se procederá a colocar los resultados debidos a cada una de las conversiones.

Las capturas corresponden al momento del pico de trabajo más alto detectado durante la conversión del texto.

También se pueden observar los consumos referidos a los procesos encargados de la creación de nuevos ficheros en el sistema.

Nombre:	Fichero01.pdf	Tamaño:	2.06 MB

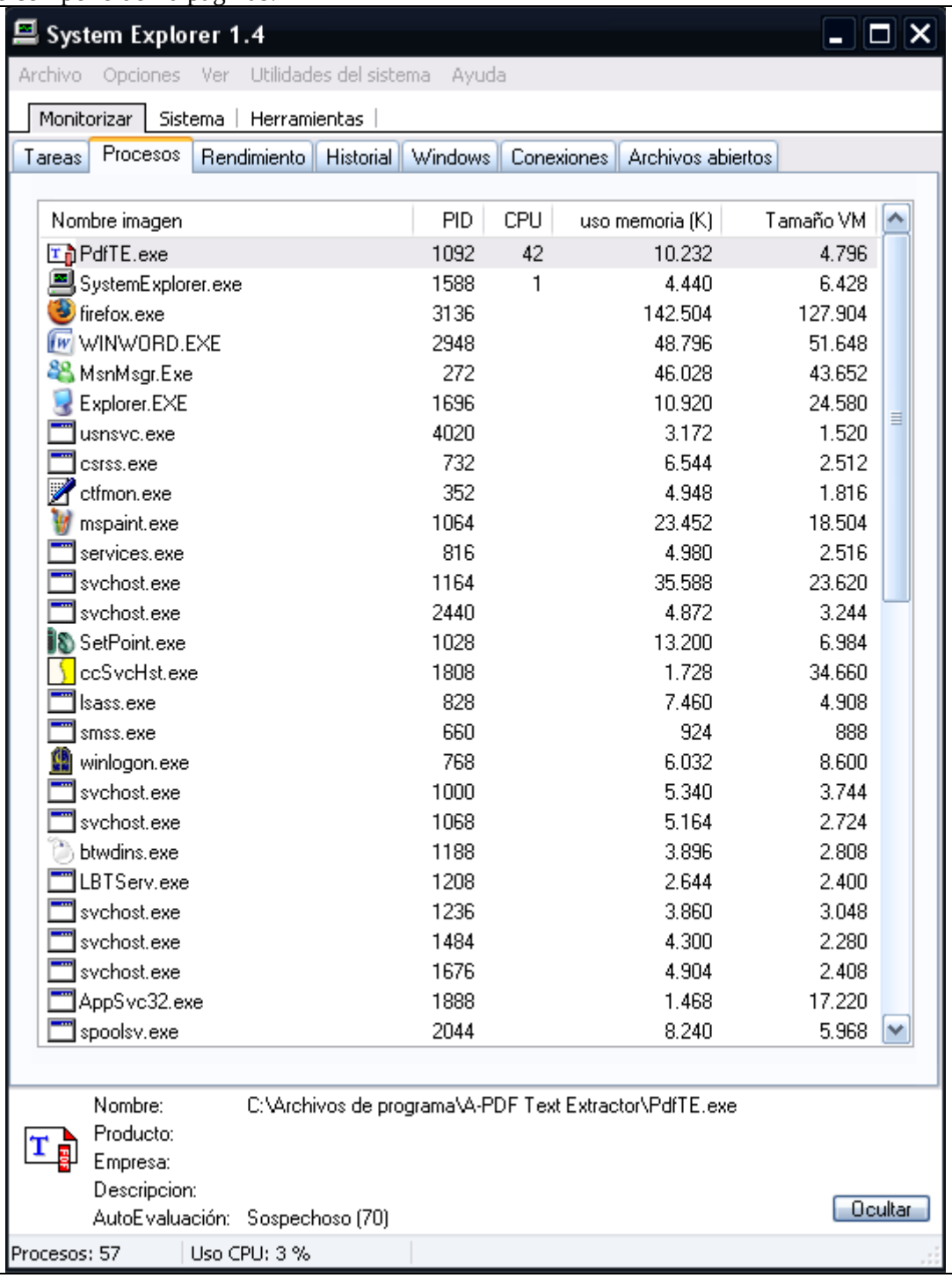
Este fue el fichero de referencia para la primera prueba.
Se puede considerar un fichero tipo, presenta el Abstract a una columna y el resto del fichero a dos columnas.
Incluye imágenes de un tamaño pequeño.
Se compone de 9 páginas.



En comparación frente al XPDF se puede ver que el consumo total de CPU en el momento de guardado es menor pero el consumo de memoria RAM es más alto, el tamaño del fichero se puede observar en la carga en la RAM.

Nombre:	Fichero02.pdf	Tamaño:	1.11 MB

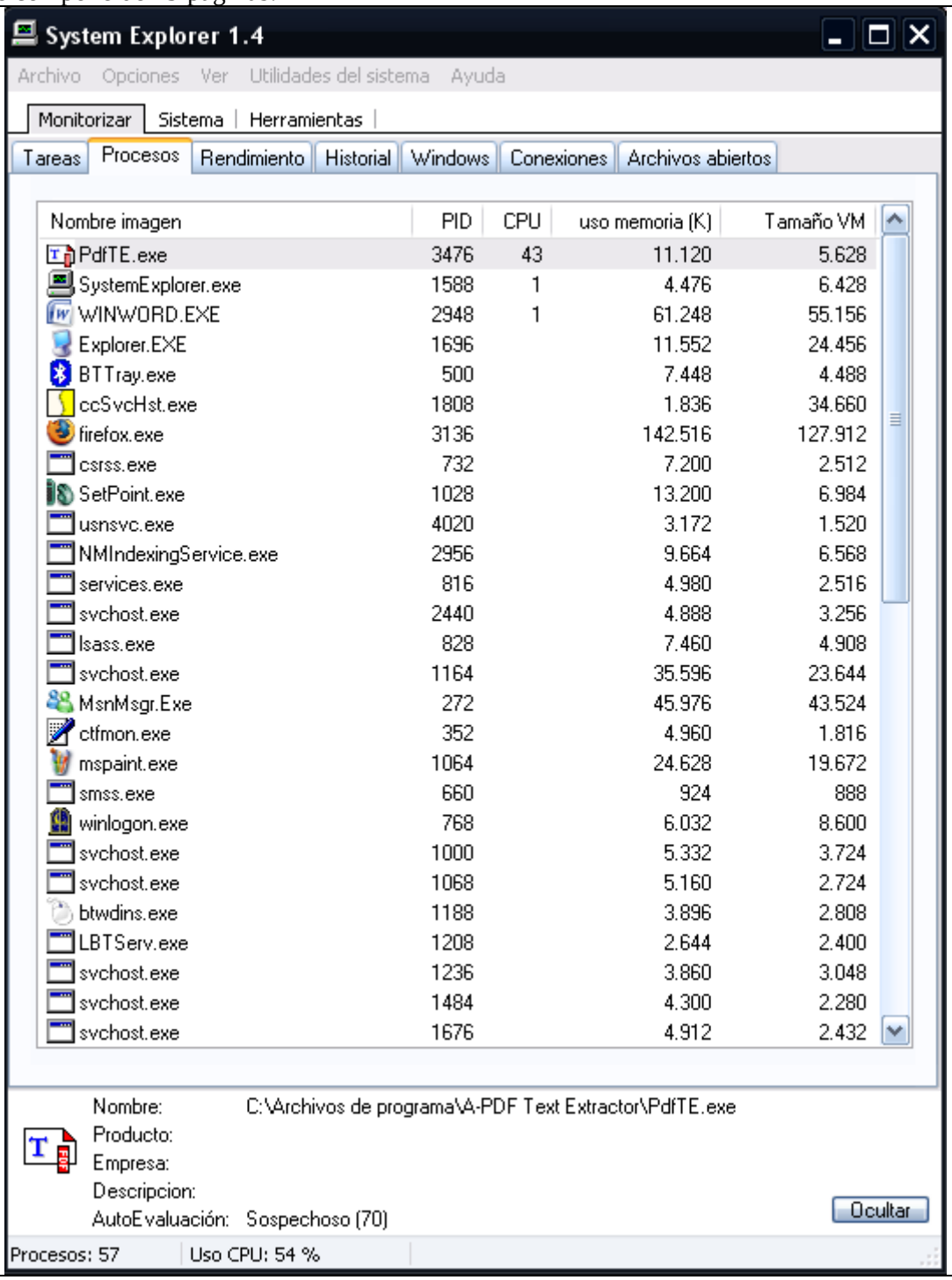
Se puede considerar un fichero tipo, presenta Abstract pero no identificado como tal a una columna y el resto del fichero a dos columnas.
Incluye imágenes de tamaños variables.
Se compone de 10 páginas.



Comparando con el resultado anterior se puede ver que el consumo de CPU también es más reducido y el de RAM es más bajo al tratarse de un fichero más ligero.

Nombre:	Fichero03.pdf	Tamaño:	178 KB

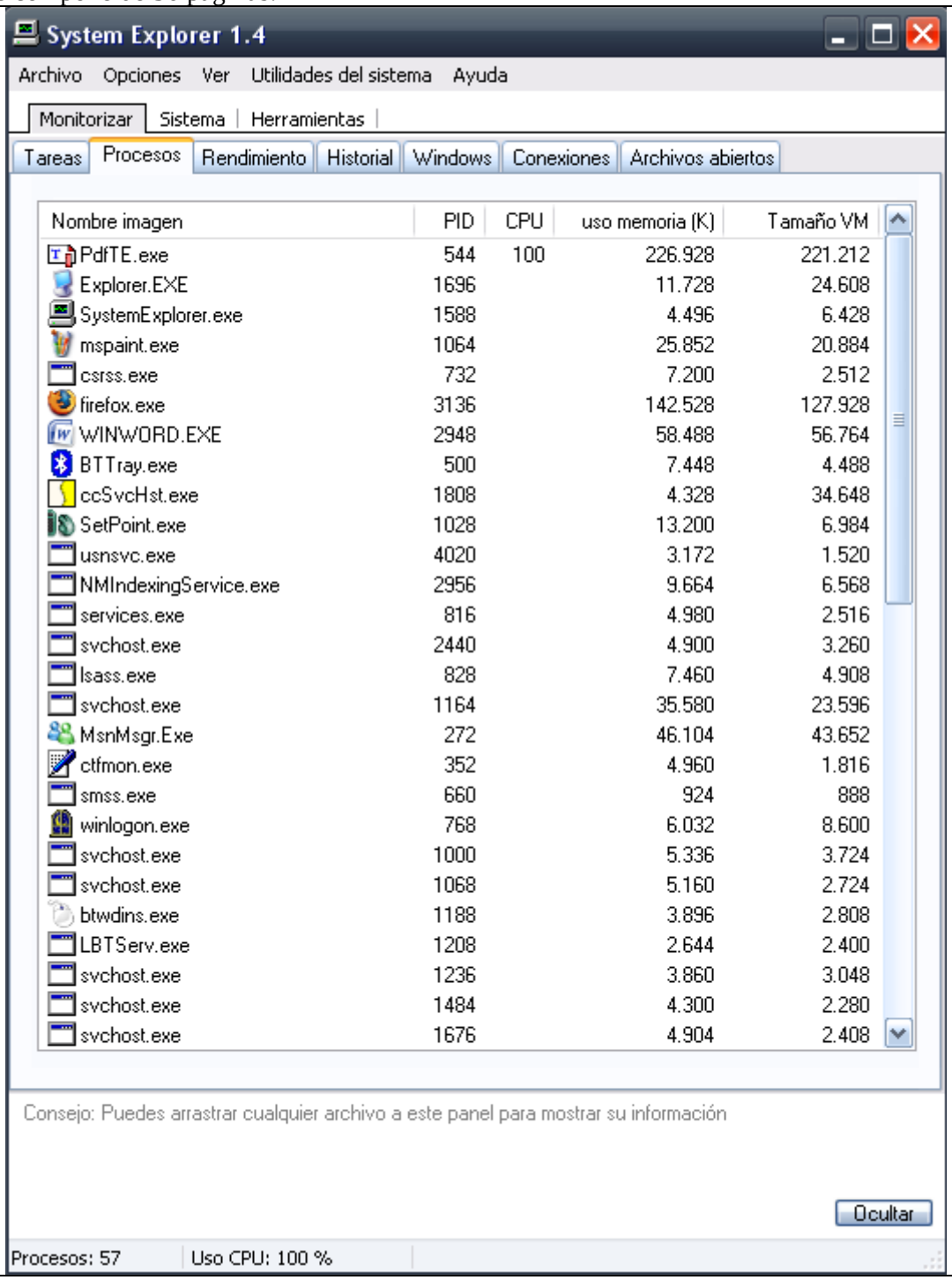
Se trata de un artículo aun no completo no acorde al formato estándar.
Contiene todas las imágenes al final del texto.
Se compone de 19 páginas.



Para este fichero de tamaño reducido y página del Abstract de gran sencillez se observa un consumo mucho más bajo tanto de CPU en cuanto al consumo de RAM se mueve dentro de unos valores normales.

Nombre:	Fichero11.pdf	Tamaño:	4.92 MB

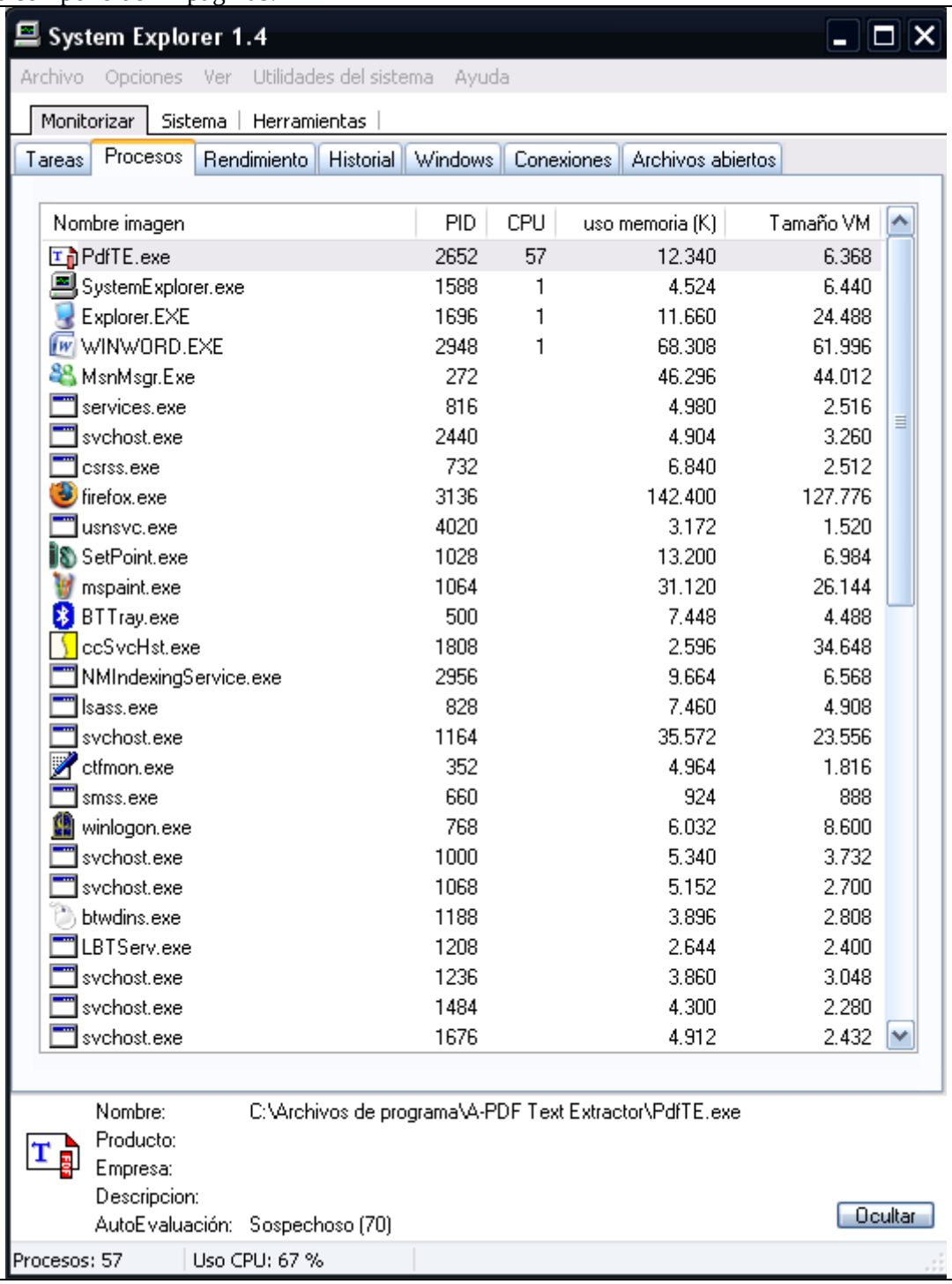
Se trata de un artículo aun no completo no acorde al formato estándar.
Contiene todas las imágenes al final del texto, se tratan de unas imágenes muy ricas en datos puesto que son formas de onda y son altamente detalladas.
Se compone de 36 páginas.



Al igual que ocurría en el XPDF el consumo de CPU sube hasta copar el 100% del procesador lo que sí que podemos observar es como se dispara el consumo de la memoria RAM, que como ya se ha indicado es debido a la gran resolución de la imagen presente.

Nombre:	Fichero13.pdf	Tamaño:	230 KB

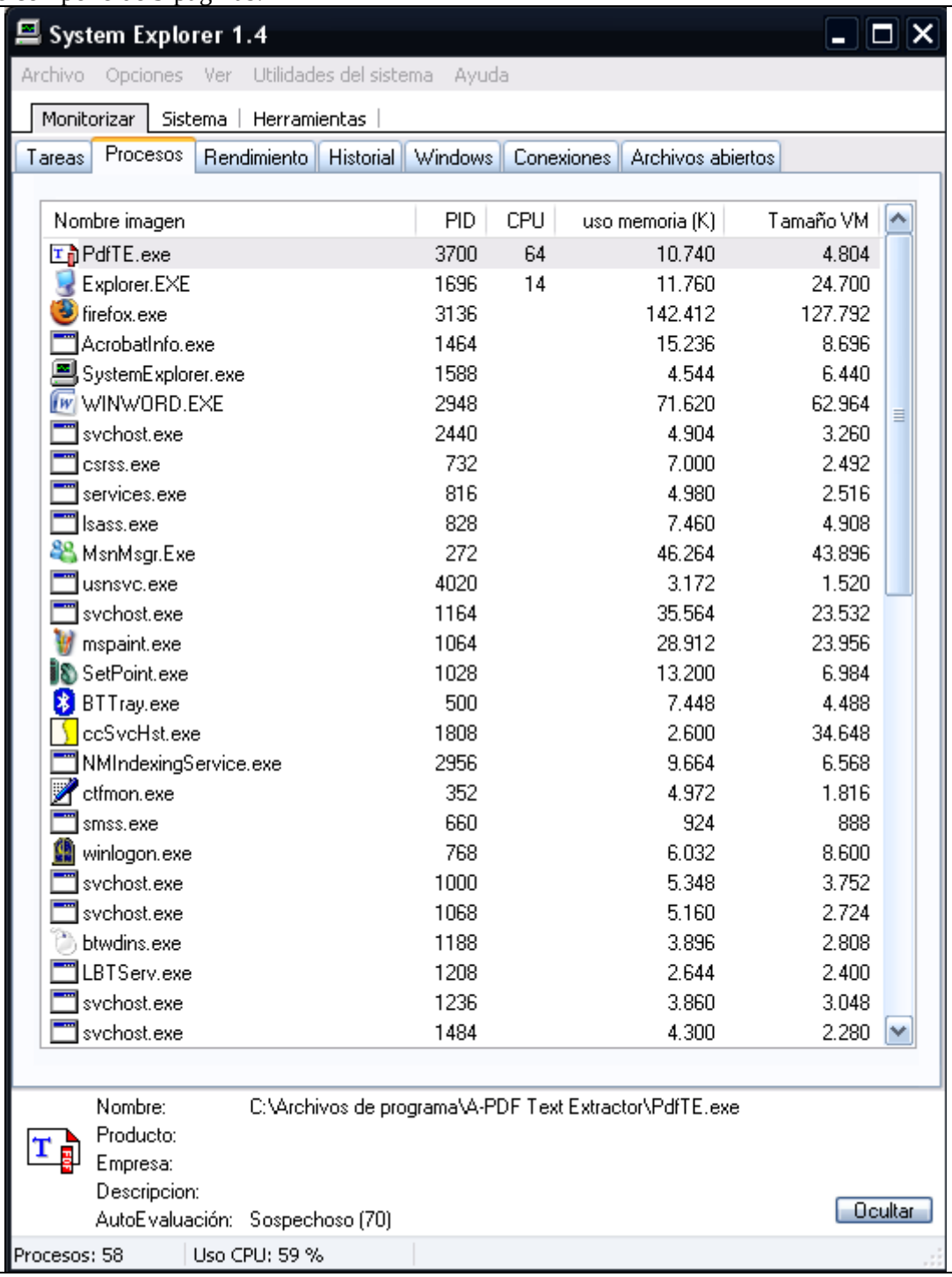
Se trata de un artículo aun no completo no acorde al formato estándar.
Contiene todas las imágenes al final del texto, se tratan de unas imágenes muy simples, son unos esquemáticos también presenta tablas de datos.
Se compone de 17 páginas.



Fichero simple y sin ninguna complicación, presenta un consumo de recursos más el elevado que el XPDF podemos empezar a deducir que se las imágenes no es el punto fuerte de este programa.

Nombre:	Fichero16.pdf	Tamaño:	793 KB

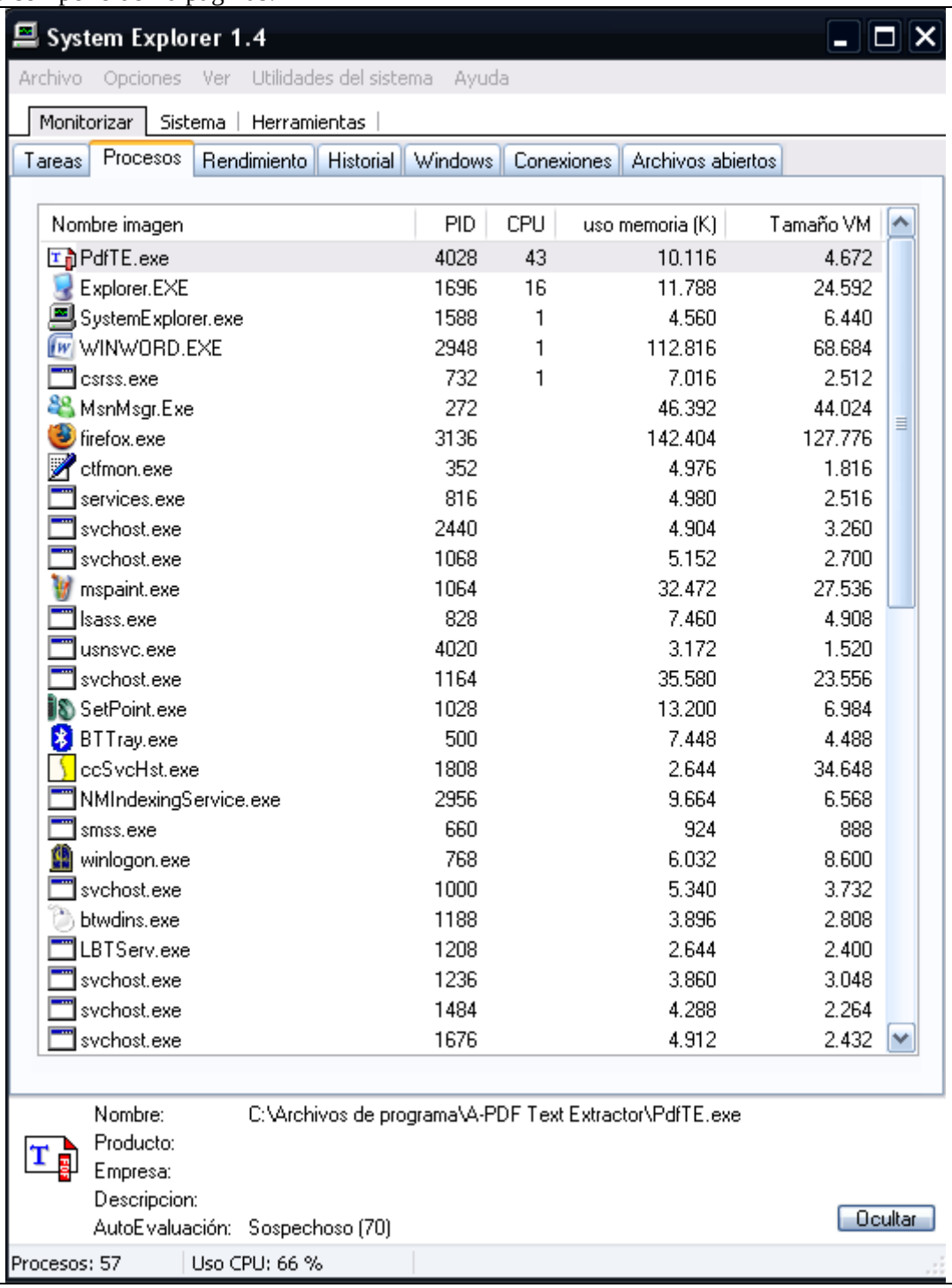
Se puede considerar un fichero tipo, presenta el Abstract a una columna y el resto del fichero a dos columnas.
Incluye imágenes de diverso tamaño una de ellas especialmente rica.
Se compone de 9 páginas.



Este fichero también presentaba complicaciones para el XPDF, la presencia de imágenes ricas vuelve a repercutir en uso más alto de CPU que el XPDF, en este caso no se ve descontrolado el crecimiento en memoria RAM.

Nombre:	Fichero20.pdf	Tamaño:	5.74 MB

Se puede considerar un fichero tipo, presenta el Abstract a una columna y el resto del fichero a dos columnas.
Incluye imágenes de diverso tamaño son medianamente ricas en cuanto a detalles, se prevé que el programa presente carga de trabajo.
Se compone de 10 páginas.



Otro fichero también complejo para el XPDF y aquí se vuelve a presenciar lo mismo si aparecen imágenes el consumo de CPU se verá alterado, y el de memoria RAM solo en caso en que estas sean de una gran resolución o riqueza.

Tabla Resumen:	
Fichero01.pdf	47%
	23,82MB
Fichero02.pdf	42%
	10,232 MB
Fichero03.pdf	43%
	11,120 MB
Fichero11.pdf	100%
	226,928 MB
Fichero13.pdf	57%
	12,340 MB
Fichero16.pdf	78%
	22,5 MB
Fichero20.pdf	59%
	21,904 MB

A la vista de estos resultados obtenidos se puede concluir las siguientes afirmaciones:

La extensión en páginas en el fichero no es un motivo de consumo suficientemente significativo, como es de pensar cuantas más páginas tiene el fichero más tiempo le costara la conversión.

El peso de los ficheros se nota sensiblemente en la carga del programa, puesto que lo que hace es una carga total del fichero al principio.

En cuanto al uso de memoria RAM se ha visto que el programa hace un uso bastante intensivo de memoria frente a la poca carga que provoca el XPDF no la hace.

Concluyendo, el programa presenta un comportamiento bueno aunque en determinadas ocasiones se vuelve un gran consumidor de recursos, convirtiendo las paginas es sensiblemente más lento como se podrá ver luego en el test de productividad, es un gran programa a tener en cuenta.

2ª Parte:

Estudio de productividad.

Resultados de la única estrategia disponible:

Tiempo: 2 Minutos y 58 Segundos lo que es un total de 178 segundos.
El número de tareas es de 27.

Total de productividad= 178 segundos /27 Tareas = 6.59 segundos por tarea.

Resultados de conversión en consola:
No disponible.

Suponiendo que tuviésemos un trabajador ideal sin fatiga trabajando 8 horas en exclusivo a esta tarea, lo que hace un total de 28.800, segundos trabajando de la manera únicamente evaluada obtendríamos que convertiría 4.370 ficheros en esas 8 horas.

2.3.3 OTROS ASPECTOS.

En este punto se tendrán en cuenta otros aspectos que consideramos importantes en estos programas, como son el comportamiento en la conversión del pie de figuras, las tablas y los caracteres especiales. Todos muy frecuentes en publicaciones científicas.

A la hora de escoger los programas para evaluar más en profundidad, se ha tenido en cuenta también la calidad de la conversión. Uno de esos aspectos que diferencian los resultados entre distintos programas son esos detalles. El XPDF especialmente ha mostrado unos resultados muy satisfactorios.

En la conversión de las tablas depende si se mantiene el layout del archivo original, en ese caso el resultado sería el contenido de la tabla tabulado para mantener el formato original. En el caso de no mantener el layout, por defecto en el APDF y opcional en el XPDF, se extrae el texto de la tabla y se deja de forma secuencial.

Para los caracteres especiales, el XPDF permite descargar de forma opcional los archivos de distintos idiomas. Entre ellos los caracteres del alfabeto griego muy utilizados en artículos de carácter científico. Por defecto, la conversión se hace en UTF-8 por lo que no es necesario especificarlo. El A-PDF utiliza el estándar ISO-8859-1 por lo que daría problemas para convertir caracteres especiales.

Las figuras al igual que con las tablas depende si se mantiene el layout o no. En el caso de que se mantenga se dejaría el hueco de la figura y debajo escribe el pie de foto. Si no se mantiene el layout el texto lo inserta de forma secuencial en el lugar correspondiente. Hemos notado un comportamiento extraño con los pies de figuras y es que los textos que están en negrita se duplican al convertirlo a texto, esto ocurre en todos los programas que hemos analizado por lo que aparentemente es más un problema en la especificación del formato PDF.

3. RECOMENDACIÓN FINAL

A la luz de los resultados de la última fase del estudio donde se han tenido en cuenta los aspectos de consumo de recursos hardware de estos dos programas XPDF y A-PDF Text Extractor y de la productividad obtenida, nos vemos en la decisión de recomendar como programa que mejor se comporta el XPDF por los siguientes motivos.

1 Aunque presente un uso más intensivo de la CPU el tiempo que la ocupa es mínimo lo cual no repercute excesivamente en el rendimiento de la maquina.

2 El consumo en memoria RAM no es excesivamente alto en ambos programas el A-PDF Text Extractor se ha descontrolado en una de las conversiones, lo cual podría causar problemas si andamos escasos de recursos maquina.

3 En cuanto a la productividad se puede ver que aun usando el sistema de conversión serie que es sensiblemente más lento y por lo tanto menos productivo resulta que es incluso más rápido que el uso del A-PDF Text Extractor.

También se han tenido en cuenta los aspectos de calidad y conversión especificados en el 2.2.3 donde se tenían en cuenta cosas como la conversión de caracteres especiales… se ha visto que el XPDF se comporta bien en este aspecto y por el contrario el A-PDF no entiende estos caracteres y los convierte en símbolos extraños, porque como ya hemos visto la codificación de caracteres que utiliza es la ISO-8859-1.

A la luz de estas últimas conclusiones recomendamos el XPDF como el mejor conversor, extractor en el momento aunque tenga algunos pliegues que pulir, pero al tratarse de un producto opensource estos pliegues los podríamos realizar nosotros.

4. CONLUSIONES FINALES

En un primer momento pensamos que la realización de un trabajo de consultoría tecnológica, nos resultaría bastante sencillo de realizar puesto que a priori se presenta más sencillo que un proyecto de desarrollo.

Pero estábamos equivocados puesto que uno de los mayores problemas que hemos encontrado es la falta de experiencia en la realización de este tipo de proyectos ya que durante el transcurso de la titulación en ningún momento nos encontramos ante este tipo de trabajos.

Durante bastante tiempo estuvimos discutiendo acerca de cómo realizar y acometer esta tarea, porque teníamos bastante desacuerdo de criterios puesto que para lo que uno era interesante para el otro no lo era y así pasaba también en la otra dirección.
Cuando ya conseguimos ponernos de acuerdo nos tocó la fase de documentación, de la cual aquí se ha dicho poco puesto que en esta rama aún queda mucho trabajo por hacer como hemos podido ver gracias a nuestra propia experiencia.
Una vez solventados estos 2 problemas iníciales la fase de experimentación resulto más interesante de lo esperado puesto que cada nueva herramienta aportaba alguna dificultad o algún reto que las otras no aportaban, por lo cual el proyecto se volvía mas interesante por momentos.
Y por último la fase en la que nos encontramos ahora, redacción de resultados y conclusiones del proyecto, podemos decir que ha sido una experiencia interesante este nuevo sistema de trabajo, el que se podía llamar consultor técnico.

En resumen la realización del proyecto ha resultado agradable para nosotros y nos ha supuesto el adquirir una nueva dinámica de trabajo para futuros trabajos de este tipo que se nos puedan presentar.

5. PERSPECTIVAS DE FUTURO

Teniendo como objetivo final la conversión de las publicaciones al formato PubMed Central DTD, destacaríamos una serie de recomendaciones para implementar una aplicación que se encargara de esta tarea.

Como hemos visto en los resultados de las distintas pruebas, el programa en el que basarse sería el XPDF, ya que por el tipo de licencia es posible modificar el código del programa a gusto del cliente. La calidad de los resultados es la mejor de las aplicaciones testeadas y además tiene una gran cantidad de opciones de configuración.

Una de las primeras tareas en el desarrollo de la aplicación sería eliminar los pequeños defectos que hemos encontrado en la conversión, tales como la extracción de cabeceras y pies de página o la duplicación de los pies de figura. Ya que añaden información no necesaria para el postprocesado del texto.

Uno de los mayores problemas que se presentan sería identificar las diferentes secciones del documento ya que no todas las publicaciones mantienen el mismo formato. Por ejemplo en las diversas publicaciones que hemos trabajado, una de ellas identifica correctamente cada sección aunque la mayoría de ellas presenta el texto de una forma continua, lo cual hace muy complicado saber en que punto del artículo te encuentras.

La forma de acometer este problema podría ser el uso de inteligencia artificial, usando bases de conocimiento basadas en los formatos de un conjunto de publicaciones e intenta inferir de esto en que sección del documento nos encontramos. Tarea bastante complicada ya que los textos tratan de temas muy diversos.

6. BIBLIOGRAFIA.

[1] XPDF http://www.foolabs.com/xpdf/home.html

[2] Cool PDF READER http://www.pdf2exe.com/reader.html

[3] Easy PDF to Text converter http://www.pdf-to-html-word.com/pdf-to-text/

[4] A-PDF Text Extractor http://www.a-pdf.com/text/

[5] PDF2TEXT http://www.traction-software.co.uk/PDF2text/index.html

[6] LD-Getter http://www.pdftoall.com/ld-getter.htm

[7] LD-Getter Pro http://www.pdftoall.com/

[8] PDF Plain Text Extractor http://www.retsinasoftware.com/extract-convert-pdf-to-text.htm

ANEXO

BioMed Central

Open Access

BMC Developmental Biology

Research article

A genome-wide *in situ* hybridization map of RNA-binding proteins reveals anatomically restricted expression in the developing mouse brain

Adrienne E McKee[†1,2], Emmanuel Minet[†2,3], Charlene Stern[2], Shervin Riahi[2], Charles D Stiles[2,4] and Pamela A Silver*[1,2]

Address: [1]Department of Systems Biology, Harvard Medical School, Boston, MA 02115 USA, [2]Department of Cancer Biology, The Dana-Farber Cancer Institute, Boston, MA 02115 USA, [3]URBC-FUNDP, 61 rue de Bruxelles, 5000 Namur, Belgium and [4]Department of Microbiology and Molecular Genetics, Harvard Medical School, Boston, MA 02115 USA

Email: Adrienne E McKee -adrienne_mckee@student.hms.harvard.edu; Emmanuel Minet -emmanuel.minet@fundp.ac.be; Charlene Stern -cstern@foleyhoag.com; Shervin Riahi -shervin@gmail.com; Charles D Stiles -charles_stiles@dfci.harvard.edu; Pamela A Silver* -pamela_silver@dfci.harvard.edu

* Corresponding author †Equal contributors

Received: 06 May 2005 Accepted: 20 July 2005

BMC Developmental Biology 2005, **5**:14 doi:10.1186/1471-213X-5-14

This article is available from: http://www.biomedcentral.com/1471-213X/5/14

Abstract Background: In eukaryotic cells, RNA-binding proteins (RBPs) contribute to gene expression by regulating the form, abundance, and stability of both coding and non-coding RNA. In the vertebrate brain, RBPs account for many distinctive features of RNA processing such as activity-dependent transcript localization and localized protein synthesis. Several RBPs with activities that are important for the proper function of adult brain have been identified, but how many RBPs exist and where these genes are expressed in the developing brain is uncharacterized.

Results: Here we describe a comprehensive catalogue of the unique RBPs encoded in the mouse genome and provide an online database of RBP expression in developing brain. We identified 380 putative RBPs in the mouse genome. Using *in situ* hybridization, we visualized the expression of 323 of these RBP genes in the brains of developing mice at embryonic day 13.5, when critical fate choice decisions are made and at P0, when major structural components of the adult brain are apparent. We demonstrate i) that 16 of the 323 RBPs examined show neural-specific expression at the stages we examined, and ii) that a far larger subset (221) shows regionally restricted expression in the brain. Of the regionally restricted RBPs, we describe one group that is preferentially expressed in the E13.5 ventricular areas and a second group that shows spatially restricted expression in post-mitotic regions of the embryonic brain. Additionally, we find a subset of RBPs that share the same complex pattern of expression, in proliferating regions of the embryonic and postnatal NS and peripheral tissues.

Conclusion: Our data show that, in contrast to their proposed ubiquitous involvement in gene regulation, most RBPs are not uniformly expressed. Here we demonstrate the region-specific expression of RBPs in proliferating vs. post-mitotic brain regions as well as cell-type-specific RBP expression. We identify uncharacterized RBPs that exhibit neural-specific expression as well as novel RBPs that show expression in non-neural tissues. The data presented here and in an online

database provide a visual filter for the functional analysis of individual RBPs.

Background

The ordered production and differentiation of cell types that occurs during nervous system (NS) development relies upon tightly regulated gene expression. In neural cells, spatial and temporal gene regulation occurs through both transcriptional and post-transcriptional mechanisms. While the transcriptional networks that direct neural cell fate and govern cell shape, position, and connectivity have been well studied [1-3], the post-transcriptional influences on neural development and gene expression are less well understood.

At the core of post-transcriptional gene regulation are RNA-binding proteins (RBPs). Proteins containing canonical RNA-binding domains (RBDs) are involved in numerous steps of nuclear and cytoplasmic RNA processing [4]. Through mRNA capping, splicing, editing, polyadenylation and nonsense-mediated decay, RBPs modulate the diversity of transcribed genes [4-6]. RBPs also affect the processing of non-coding RNAs [7]. Specific RBPs additionally enable asymmetric RNA distribution and translational regulation [8-10], two phenomena that are critical for affecting localized protein synthesis [11,12].

The importance of post-transcriptional processing in NS gene regulation is underscored by functional examples of specific RBPs [13,14]. For instance, the neuronal-specific factor Nova-1 regulates splicing of pre-mRNAs that encode components of inhibitory synapses [15]. Mice lacking Nova-1 die postnatally due to aberrant regulation of apoptotic neuronal death [16]. As a second example, RBPs encoded by the quaking and Musashi loci promote glial cell fate [17] and CNS stem cell self-renewal [18] by stabilizing transcripts involved in cell differentiation. Thirdly, the fragile X mental retardation protein, members of the ELAV/Hu protein family, and the Staufen proteins are involved in targeting and translational regulation of dendritic transcripts [19-21]. Additionally, the finding that long-term memory requires *de novo* protein synthesis highlights the significance of post-transcriptional processes in neural function [22,23].

Despite our knowledge of several key RBPs, much of the understanding of RBPs in the brain comes from studies of adult animals or neural cell lines. Thus, how the functional class of RBPs contributes to the positioning, growth, and diversification of cells in the developing brain is not well understood. One step towards increasing our understanding RBPs is to resolve where they are expressed. Here, we utilize the approach of *in situ* hybridization mapping [24-26] to investigate the expression of 323 RBPs within the developing mouse brain. Two stages of development were characterized, embryonic day 13.5 (E13.5), when critical cellular fate choice decisions are made and postnatal day 0 (P0), when the major structural components of the brain are apparent. We find that, in contrast to their proposed ubiquitous involvement in gene regulation, most RBPs are not uniformly expressed. The majority of RBPs profiled demonstrates spatially restricted expression in the brain or in other peripheral tissues examined. The data presented here and in an online database afford a visual filter for the functional analysis of individual RBPs in the developing mammalian NS.

Results

Mouse RBPs were identified according to gene sequence

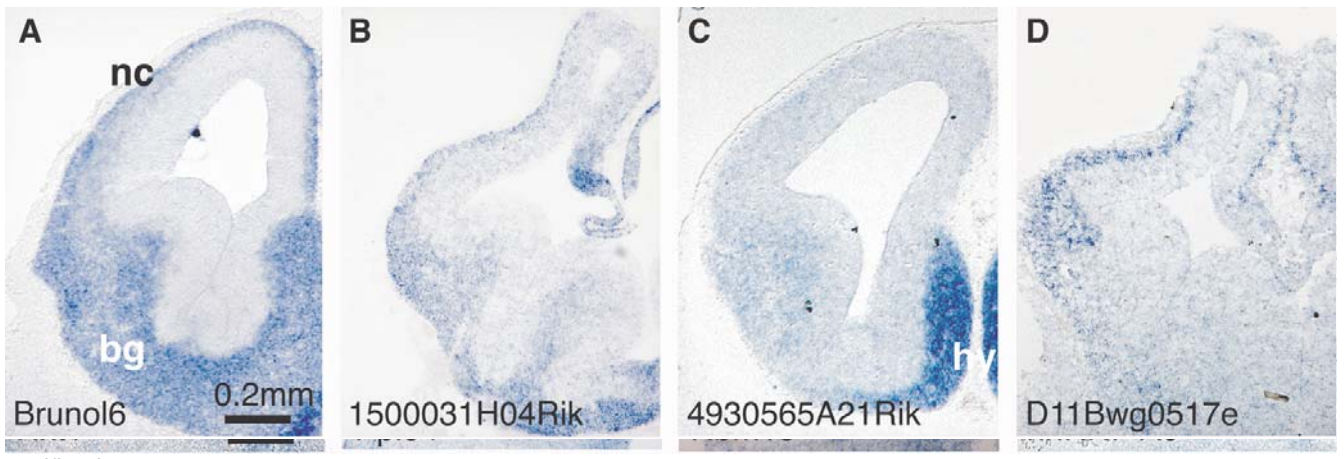
The RNA recognition motif (RRM), the hnRNP K-homology (KH) domain, and the double-stranded RNA-binding domain (dsRM) are evolutionarily conserved, well-characterized domains known to bind either single or double-stranded RNA [27-29]. Sequence similarity searches and structural analyses of these domains have led to the ability to predict other RBPs based on primary coding sequence [29]. To identify unique genomic loci that encode putative RBPs in the mouse genome, we analyzed existing public [30,31] and private [32] databases for sequences containing one or more RBD. Candidates were classified as RBPs only if their predicted protein sequence contained a Protein Families Database (Pfam)-defined RBD [31].

We identified 290 genes harboring one or more RRM, KH, or dsRM sequences. We also identified 32 genes encoding other domains shown to interact with RNA, including the zinc knuckle, G-patch, PIWI, DEAD box RNA helicase, and TUDOR domains. Finally, as the absence of a canonical RBD does not preclude interaction with RNA, we sought 58 additional genes known or predicted to be associated with RNA processing. In total, this collection contains 380 putative RBPs. Additional file 1 lists the number of genes, per RBD, identified and analyzed by *in situ* hybridization. A list of all genes and primer sequences is given in Additional file 2.

RBP expression in the developing mouse brain was analyzed by in situ *hybridization*

To localize RBP expression, we preformed *in situ* hybridization on whole head tissue sections of E13.5 embryos and P0 mice. We designed gene-specific primers to produce 400–700 bp probes for 340 candidate RBPs. These primer sets were used to perform PCR on cDNA prepared from embryonic or P0 mouse brains. A small number of probes were obtained from mouse intestine, liver, kidney, or testes cDNA. 323 genes (95%) showed positive PCR products (data not shown). Following subcloning, antisense digoxygenin-labeled riboprobes were prepared and hybridized against coronal head and transverse upper-body sections (to include the brain and spinal cord, respectively). Digital images of the entire *in situ* hybridization set have been deposited in the Mahoney RNA-Binding Protein Expression Database [33].

Figure 1 RBP expression in proliferative zones of the E13.5 mouse forebrain. *In situ* hybridization patterns for four RBPs on sections through the forebrain of E13.5 mice. Labels indicate Locuslink gene names. All images show the same magnification.



RBPs exhibit restricted expression in the developing mouse brain

Several neural-specific RBPs have been identified, yet how many others demonstrate this degree of specificity is unknown. Of the genes examined we found 16 RBPs (listed in Additional file 2) that exhibit NS-restricted expression in the tissues analyzed. Among this list are known examples of neuronal-specific RBPs including Nova-1 [34], the ELAV/Hu proteins B, C, and D [35], and Ataxin 2 binding protein 1 (A2bp1) [36] but additionally include putative RBPs for which expression has not been reported. With the exception of one gene that was only detected at E13.5, all (15/16) of these RBPs appear brain or NS-specific at both developmental stages in the tissues analyzed. Overall, these RBP encoding genes are not limited in expression to one brain region but are found in multiple brain or NS structures.

RBPs show spatially restricted expression in anatomically distinct brain regions

We find that greater than half of the RBPs profiled exhibit spatially restricted expression. Of the 323 genes examined, 221 demonstrate localized, enriched expression in one or more discrete brain regions in addition to detectable expression in non-NS tissues. We divided the E13.5 and P0 CNS into five and eight general areas for annotation, respectively: the E13.5 precortical area, the striatum (and other basal ganglia), the periventrical areas, hindbrain, and spinal cord, as well as the P0 cortex, striatum, hippocampus, thalamus, hypothalamus, midbrain, hindbrain, and spinal cord. The presence or absence of expression for each RBP was analyzed visually at each location and is annotated in Additional file 3. Very few of the 221 RBPs with spatially restricted expression patterns were expressed in only one brain region, however most (73%) showed restricted expression at both developmental stages (Additional file 3).

We observe multiple RBPs that demonstrate region-specific expression in the E13.5 ventricular areas. Shown in Figure 1 are representative RBP genes that are transcribed in mitotically-active cells in the neuroepithelia of the developing telencephalon. Among the RBPs expressed in this region occupied by neural progenitor cells, we find examples of mRNA export factors in addition to putative splicing factors and transcriptional regulators (Fig. 1). In all instances, expression in the embryonic lateral ventricular zone is accompanied by expression in the periventricular areas of the 3rd and 4th E13.5 ventricles and often by heightened expression in the P0 subventricular zone [33]. Notably, we observed this pattern of expression for the dsRM-containing Musashi proteins [33]. Our results are consistent with the documented expression of Msi1 and Msi2 [37,38].

Multiple RBPs show restricted expression in post-mitotic regions of embryonic brain. Presented in Figure 2 are examples of four putative RBPs that demonstrate region-specific expression in areas containing post-mitotic neurons. Transcripts of the genes encoding the RRM protein Brunol6 and the predicted zinc-knuckle protein 1500031H04Rik appear pan-neuronal at both developmental stages (Fig. 2A, 2B and [33]). Expression of the RRM-containing RIKEN gene 4930565A21 is most pronounced in the ventral telencephalon, while D11Bwg0517e is found in the precortical layer, the thalamic area and hindbrain (Fig. 2C, 2D and [33]). Among the genes that occupy post-mitotic regions of the developing brain we additionally observe members of the ELAV/Hu family as well as other RBPs that have well-documented neuronal expression [34,35].

Figure 2 RBP expression in post-mitotic areas of the E13.5 mouse forebrain. *In situ* hybridization patterns for four RBPs on sections through the forebrain of E13.5 mice. Labels indicate Locuslink gene names. bg, basal ganglia; hy, hypothalamus; nc, neocortex. All images show the same magnification.

RBPs demonstrate cell-type specific expression in the P0 mouse retina

As our *in situ* hybridization analyses were performed on sections through whole head, we were able to visualize RBP expression in the developing retina. The vertebrate retina provides a distinctive system for studying CNS development as its seven major neural cell types are readily distinguished from one another by their morphology and laminar position [39]. Shown in Figure 3 are examples of the diversity of RBP expression in the P0 retina. The RRM-containing A2bp1 is expressed in the retinal ganglion cell layer (GCL), which contains primarily retinal ganglion cells and a small number of displaced amacrine cells (Fig 3A, 3B). The KH-domain encoding gene poly(rC) binding protein 3 (Pcbp3) shows dramatically enriched expression in the inner nuclear layer (INL) (Fig. 3C and 3D), possibly indicating localization to the bipolar neuron cell bodies that occupy the scleral portion of the INL. Notably, both A2bp1 and Pcbp3 show restricted expression in post-mitotic regions of the E13.5 and P0 brain [24,36]. Transcripts of the RRM-encoding scaffold attachment factor B (Safb) and of the three-RRM containing SPOC gene Rbm15 are expressed in the outer neuroblastic layer of the retina (Fig. 3E–H). Safb, but not Rbm15, is additionally expressed in the GCL, possibly in the Müller glia. Both Safb and Rbm15 show enriched expression in neuroepithelia of the ventricular zone (Fig. 1 and [33]).

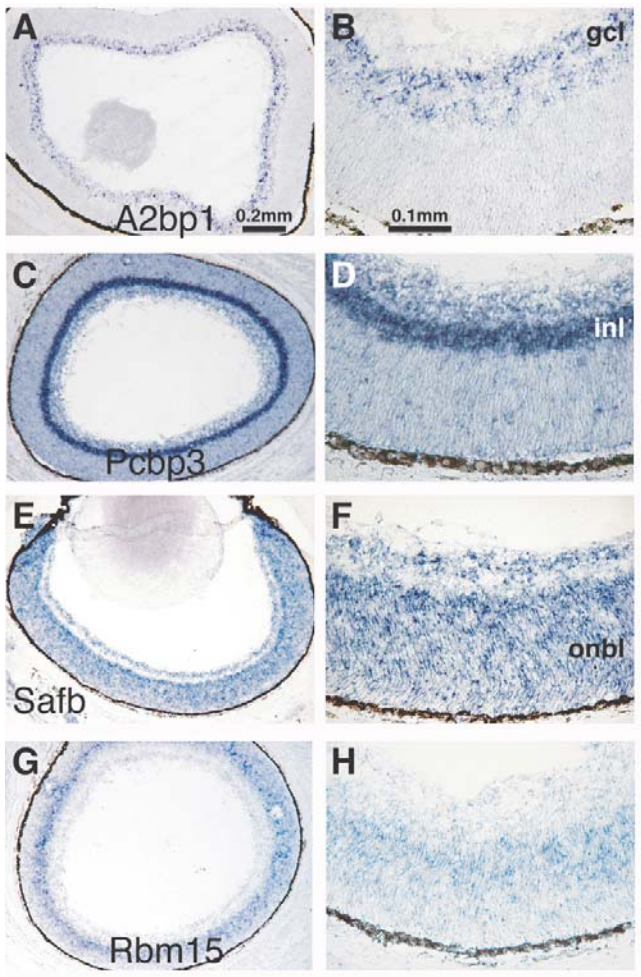


Figure 3 Diversity of RBP expression in major cellular subtypes of the P0 retina. *In situ* hybridization for four representative RBPs that exhibit laminar-specific expression in the P0 mouse retina. Labels indicate Locuslink gene names. A, B) A2bp1, C, D) Pcbp3, E, F) Safb, G, H) Rbm15. Panels A, C, E, and G show the same magnification. Panels B, D, F, and H show the same magnification. gcl, granule cell layer; inl, inner nuclear layer, onbl; outer neuroblastic layer.

A systems-based view of RBP expression

Gene regulation by RBPs is believed to occur through coordinated, combinatorial interactions with RNA. During the course of this study we identified multiple RBPs that are coordinately expressed in the brain and other tissues. We find 48 genes (listed in Additional file 4) that show elevated expression in proliferating areas of the embryonic and postnatal brain as well as in postnatal nasal epithelia, teeth, and thymus. Presented in Figure 4 are expression data for snRNP E and Son, two representative examples of this "synexpression group" of genes that share a similar, complex pattern of expression. Further examples are shown in Additional file 5. This same expression distribution has been observed for the polypyrimidine tract-binding protein, PTBP1, and our data are consistent with previous findings [40]. Notably, the protein products of many of the genes listed are understood to interact either physically or genetically.

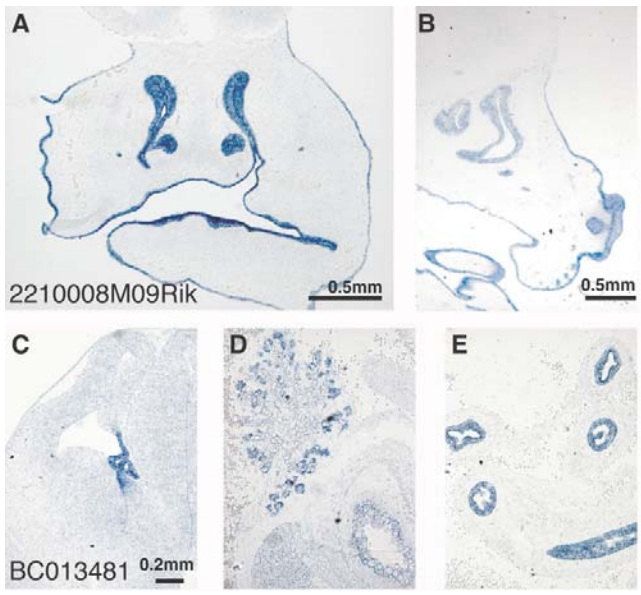
RBPs show restricted expression in non-NS tissues

As our analyses were performed on whole head and upper thoracic tissues, our data provide detailed information about RBP expression in developing cranial facial tissues. We identified putative RBPs that display tissue-restricted expression in non-NS structures (listed in Additional file 3). Figure 5 presents *in situ* hybridization results for two RRM-encoding transcripts that show highly restricted expression in different epithelial tissues. The Riken gene 2210008M09 is transcribed in epithelia covering the facial skeleton (Fig. 5A, 5B), while the gene BC013481 is expressed in the choroid plexus (Fig. 5C) and in the lining of the intestine and placenta (Fig. 5D, 5E).

Discussion

Neural cells utilize multiple forms of post-transcriptional gene regulation. While RBPs are believed to be potent modulators of post-transcriptional processes, little is known about how this functional class is expressed in the developing brain. As a first step towards increasing our knowledge of RBPs we chose to investigate the spatial and temporal expression of genes that encode motifs known

Figure 4 Representative examples of RBP synexpression in E13.5 and P0 mouse tissues. snRNP E and Son are transcribed in the perventricular areas of



the E13.5 brain (A, E), in the P0 subventricular area of the lateral ventricle (B, F), in the external granule layer of the P0 cerebellum (C, G), as well as in postnatal developing teeth (D, H).

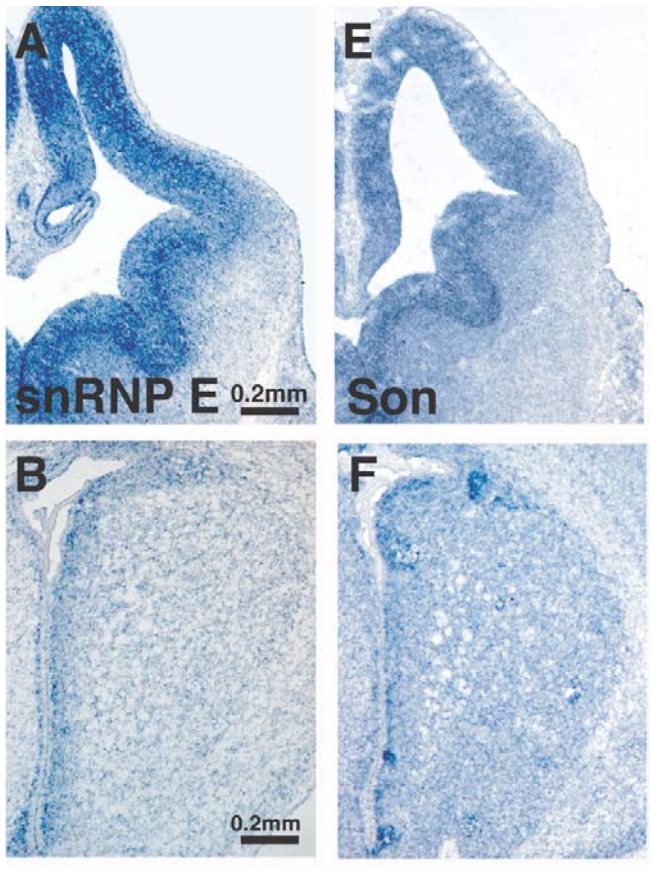
Figure 5 *In situ* hybridization profiling uncovers the non-neural, restricted expression of novel RBPs. Data from ISH performed on (A, C) coronal E13.5 and on (B, D, E) E15 sagittal sections are presented for RRM-encoding RBPS. A, B) The Riken gene 2210008M09 is transcribed in epithelia covering the facial skeleton. C-E) BC013481 is detected in the choroid plexus, in the intestinal lining, and in the lining of the placenta. Panels C-E show the same magnification.

An even greater number of RBP genes however demonstrate spatially restricted expression in distinct regions of the developing brain.

Within the CNS, most of the RBPs examined show non-uniform, heightened expression in anatomically discrete structures. Tissue differences in the expression levels of individual genes could indicate distinctive protein requirements among cell types, beyond that of tissue-specific RBPs [41]. There is precedent for differential requirements of individual RBPs, as tissue-specific RNA splicing is achieved partly through combinatorial, stoichiometric differences among splicing factors within various cells [42]. It is from this local enrichment within different cell types or tissues that we can begin to hypothesize as to the functional significance of individual genes as well as to the importance of groups of similarly expressed RBPs.

Our study has identified RBPs that display spatially restricted expression in distinct regions of the developing mouse brain. One set of RBPs (Fig. 1) is found in the E13.5 ventricular areas. A second set demonstrates spatially restricted expression in post-mitotic regions of E13.5
brain (Fig. 2). Based on their pattern of expression, these RBPs may have roles in neural proliferation, cell fate choice and cell

to interact with RNA. We find a small set of RBPs that show neural-specific expression in the tissues analyzed.



migration, or in neuronal function, respectively. We also identified novel RBPs that are expressed in tissues of mesodermal and endodermal origin (Fig. 5). The highly restricted expression of these genes may indicate an explicit role for these RBPs in their respective epithelia. Additionally, the cell-type specificity RBPs found in the P0 retina (Fig. 3) illustrates the diversity of RBP expression. The specialized expression of these RBPs may be indicative of a dedicated function in the specified tissues.

By visual inspection of *in situ* hybridization data, we find a subset of RBPs that are coordinately expressed in multiple tissue types. These genes display heightened expression in the periventricular areas of the E13.5 brain and spinal cord as well as marked expression in the external granule layer of the P0 cerebellum, the lateral subventricular zones, and in teeth, nasal epithelia, and thymus (Fig. 4, Additional file 5, [33]). While not excluded from post-mitotic tissues, these RBPs are

predominately expressed in structures that are undergoing cell division.

Notably, the term 'synexpression group' has been used to describe collections of genes that function in a common process and share a similar complex spatial expression pattern in multiple tissues [43]. Among the synexpression group identified here we find examples of RBPs that are known to interact either physically or genetically (Additional file 4). For example, PTBP1 binds the splicing factors PSF [44] and hnRNP L [45] while SF2/ASF and hnRNP A1 select for 5' exon or exclusion or inclusion, respectively [46]. Our data provide visual support to a growing body of evidence that functionally-related transcripts are post-transcriptionally co-regulated [47].

Although the significance of certain splicing and mRNA export factor enrichment in proliferating regions is not known, data from multiple studies point to a role for RBPs in cell proliferation. During hippocampal development expression levels of RBPs were found to be high and then to dramatically decrease, as neurons transition from a proliferating to a post-mitotic state [48]. A number of RBPs were also identified as highly expressed in a molecular characterization of gastric epithelial progenitor cells [49,50]. Furthermore, protein levels of hnRNPs and snRNPs were found to be down-regulated upon stimulated growth inhibition of myeloid cells [51]. Therefore, it is likely that a role for RBPs during cell proliferation and cell fate determination exists in multiple tissue types.

Conclusion

In summary, the data presented here provide new insight into how a distinct functional gene class is expressed in the developing NS. We find that RBPs demonstrate region-specific as well as cell-type specific expression. In addition, we find that specific, proliferating regions of the embryonic and postnatal NS and peripheral tissues are similar in the expression of certain RBPs. These data serve as a starting point for functional investigations into the roles of RBPs in neural development and physiology.

Methods In silico *RBP identification*

Putative RBP gene sequences were identified by homology-based whole genome screening using public and private databases: Celera Panther Families, Protein Families Database (Pfam), and Genbank [30-32]. Classification as an RBP was based on the presence of one or more RRM, KH, or dsRMs, as defined by Pfam databases [31]. Databases were also mined for zinc-knuckle, G-patch, PIWI, DEAD-box helicase and Tudor domain-containing sequences and for known factors involved in mRNA splicing, editing, transport, and stability. Genes with multiple RNA-binding domains were assigned to a single sub-family. Unique gene identity was verified by LocusID numbers. As of March 1, 2004, a total of 357 unique genes were identified from these sources. An additional 26 RRM, KH, and dsRM proteins have been identified as of March 7, 2005.

PCR primer design

PCR primer pairs were designed for each identified RNA-binding protein locus. PCR primer sequences were designed with approximately 60% GC content, spanning 400–700 base pairs of primarily the gene's coding sequence. Additional primer pairs were designed for targets that did not initially yield PCR products.

Cloning

Total RNA was obtained from E13.5, P0, or adult C57/BL6 mouse brains (Charles River Laboratories) by Trizol extraction (Invitrogen). Reverse transcription was performed using Superscript II reverse transcriptase and oligo-dT (Invitrogen). PCR was performed with cDNA templates using 40 cycles, 60–65°C annealing temperature, and Platinum Taq (Invitrogen) as polymerase. For a few genes, PCR was performed with cDNA templates prepared from adult brain, kidney, gut, liver, or testis tissues. Positive PCR products were cloned into TA cloning vectors (Invitrogen) and verified by restriction digest or DNA sequencing.

Probe synthesis

Gene fragments from verified plasmids were amplified by PCR using plasmid specific primers. Digoxigenin-labeled RNA probes were made, using PCR products as template and T7 or SP6 RNA polymerases (Roche). cRNA probes were ethanol precipitated and quantified by spectrophotometry.

Tissue preparation

E13.5 embryos were directly fixed overnight in 4% para-formaldehyde (0.1M PBS). P0 mice were transcardially perfused with 4% paraformaldehyde (0.1M PBS) and postfixed overnight at 4°C. After fixation, embryos and P0 mice were transferred to 20% sucrose overnight. The head, neck, and trunk were embedded separately in OCT (Tis-sue-Tek) on dry ice and stored at -80°C. Serial cryostat sections (14 µm) were cut and mounted on Superfrost Plus slides (Fisher). Ten and twenty adjacent sets of sections were prepared from E13.5 embryos and P0 mice, respectively, and were stored at -20°C until use.

Section in situ *hybridization*

In situ hybridization was performed according to Gray et al. [25]. Following pretreatment (Proteinase K), slides were pre-hybridized for 1h at 65°C in hybridization solution (50% formamide (Ambion), 5X SSC, 0.3 mg/ml yeast tRNA (Sigma), 100 µg/ml heparin (Sigma), 1X Denhardt's (Sigma), 0.1% tween, 5 mM EDTA). P0 and E13.5 brain sections were hybridized overnight with labeled RNA probe(0.8–1.2 µg/ml) at 65°C, washed in 2X SSC at 67°C, incubated with RNase A (1 µg/ml, 2X SSC) at 37°C, washed in 0.2X SSC at 65°C, blocked in PBS with 10% lamb sera, and incubated in alkaline phosphatase labeled anti-DIG antibody (Roche) (1:2000, 10% sera) overnight. Sections were washed and color was visualized using NBT and BCIP in alkaline phosphatase buffer (100 mM Tris pH 9.5, 50 mM $MgCl_2$, 100 mM NaCl, 0.1% tween-20) containing 75 µg/ml NBT (BioRad), 600 µg/ml BCIP (Roche). Staining was stopped after visual inspection. Sections were washed, fixed in 4% paraformaldehyde, and cover-slipped in glycerol [25].

Image acquisition and RBP expression database

Images were acquired and analyzed as described [25]. Images were either scanned using a Nikon Coolscan 8000 slide scanner (4000 DPI) or digitally acquired using a Leica digital camera. Image levels have been modified in Photoshop (Adobe) for clarity. Full resolution scanned images were compressed using JPEG compression, quality 10, and have been deposited in the Mahoney RNA-Binding Protein Expression Database [33].

Authors' contributions

AEM prepared tissue samples, performed data analysis and drafted the manuscript. EM performed data analysis and both EM and SR generated reagents, tissue samples, digitized the raw data, and helped build the website. CS contributed to the design of the study and prepared tissue samples. CDS and PAS conceived of the study, participated in its design and coordination and helped prepare the manuscript. All authors read and approved of the manuscript.

Additional material

Additional File 1 ***RNA-binding proteins identified*** **in silico** ***and profiled by*** **in situ** ***hybridization.*** *List of annotated RNA-binding domains and the number of family members that were identified* in silico *and analyzed by* in situ *hybridization.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471213X-5-14-S1.xls]

Additional File 2
List of 380 genes identified as putative RBPs in the mouse genome and analyzed in this study. *Columns indicate LocusID, gene name, type of RBD, primer sequences used to isolate the target cDNA, the size of the cDNA fragment, the presence call by PCR from E13.5 and P0 brain cDNA, cloning status ('c' indicates cloned, 'u' indicates uncloned, 'small' indicates that the target gene had less than 400 bp of unique sequence, 'na' indicates that cloning was not attempted), the RNA polymerase used to generate the anti-sense riboprobe, the tissue from which the cDNA was isolated (if not from E13.5 or P0 mouse brain), and whether the gene was analyzed by* in situ *hybridization ('x' indicates yes).* Click here for file
[http://www.biomedcentral.com/content/supplementary/1471213X-5-14-S2.xls]

Additional File 3 ***Complete list of gene expression patterns for all*** **in situ** ***hybridizations performed.*** *Of the 323 RBPs examined, 221 showed restricted expression patterns in the brain. The remaining genes either show restricted expression in non-neural tissues, ubiquitous expression that is difficult to distinguish from background, or no expression. Caution is needed in interpreting the results. First, non-expression could be due to the sensitivity limit of non-radioactive* in situ *hybridization. Second, the background level of individual probes may differ. Third, some probes with high background hybridization may mask the real expression of the transcript. Fourth, we cannot rule out the possibility that some probes may show variable levels of background hybridization in different brain areas, resulting in a false positive signal. Columns* ***A-D*** *describe the LocusID, gene name, type of RBD, and number (internal Mahoney reference number). Columns* ***E and, L*** *(E13.5, P0 "****Informativity****"): "1" for restricted expression in the nervous system and "0" for either ubiquitous expression that is difficult to distinguish from background or no expression. As noted in Gray et al [25], some of the genes in the "0" category show uneven signals in different brain regions and are also annotated in the subsequent columns. Columns* ***F and M*** *(E13.5, P0 "****Specificity****"): "1" for restricted expression in neural tissues only, "2" for restricted expression in neural tissue with distinguishable expression in non-neural tissue, "3" for ubiquitous or no expression, and "4" for expression in non-neural tissues only. Columns* ***GK and N-U*** *(E13.5, P0 "****Expression****"): "2" for expression, "1" for ubiquitous expression or background, "0" for no expression.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471213X-5-14-S3.xls]

Additional File 4
RNA-binding proteins belonging to a synexpression group. *Complete list of RBPs that demonstrate a similar complex pattern of expression. Columns* ***A-D*** *describe the LocusID, gene name, type of RBD, and number (internal Mahoney reference number).*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471213X-5-14-S4.xls]

Additional File 5
Examples of RBP synexpression in E13.5 and P0 mouse tissues. *Additional examples of RBPs that share a similar pattern of expression. Shown are* in situ *hybridization results of expression in the periventricular areas of the E13.5 brain (A, E, I, M, Q), in the subventricular area of the P0 lateral ventricle (B, F, J, N, R), in the external granule layer of the P0 cerebellum (C, G, K, O, S), as well as in postnatal developing teeth (D, H, L P, T). A-D) Refbp1, E-H) hnRNP A1, I-L) PTBP1, M-P) Sfpq, QR) Hnrpl. Panels A, B, E, F, I, J, M, N, Q, R show the same magnification. Panels C, D, G, H, K, L, O, P, S, T show the same magnification.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471213X-5-14-S5.png]

Acknowledgements
We are grateful to Drs. Qiufu Ma and John Alberta for critical review of this manuscript and for assistance in this work. We thank Eric Tsung, Zhaohui Cai, and Matthew McCormack for designing the website. This work has been supported by the Bernard A. and Wendy J. Goldhirsh Foundation for Brain Tumor Research and by the Charles A. Dana Foundation. AEM is supported by an institutional training grant from the National Cancer Institute (T32CA09361). EM is funded as a FNRS Researcher through the Belgian National Research Fund and by the D. Collen Research Foundation VZW and BAEF. CS received support from the American Cancer Society (PF-02128-01-MBC).

References
1. Ross SE, Greenberg ME, Stiles CD: Basic helix-loop-helix factors in cortical development. *Neuron* 2003, **39:**13-25.
2. Wilson SW, Houart C: Early steps in the development of the forebrain. *Dev Cell* 2004, **6:**167-181.
3. Bally-Cuif L, Hammerschmidt M: Induction and patterning of neuronal development, and its connection to cell cycle control. *Curr Opin Neurobiol* 2003, **13:**16-25.
4. Dreyfuss G, Kim VN, Kataoka N: Messenger-RNA-binding proteins and the messages they carry. *Nat Rev Mol Cell Biol* 2002, **3:**195-205.
5. Lasko P: Gene regulation at the RNA layer: RNA binding proteins in intercellular signaling networks. *Sci STKE* 2003, **2003:**RE6.
6. Orphanides G, Reinberg D: A unified theory of gene expression. *Cell* 2002, **108:**439-451.
7. Cullen BR: Transcription and processing of human microRNA precursors. *Mol Cell* 2004, **16:**861-865.
8. Huang YS, Carson JH, Barbarese E, Richter JD: Facilitation of dendritic mRNA transport by CPEB. *Genes Dev* 2003, **17:**638-653.
9. Antar LN, Bassell GJ: Sunrise at the synapse: the FMRP mRNP shaping the synaptic interface. *Neuron* 2003, **37:**555-558.
10. Tang SJ, Meulemans D, Vazquez L, Colaco N, Schuman E: A role for a rat homolog of staufen in the transport of RNA to neuronal dendrites. *Neuron* 2001, **32:**463-475.
11. Martin KC: Local protein synthesis during axon guidance and synaptic plasticity. *Curr Opin Neurobiol* 2004, **14:**305-310.
12. Huang YS, Richter JD: Regulation of local mRNA translation. *Curr Opin Cell Biol* 2004, **16:**308-313.
13. Agnes F, Perron M: RNA-binding proteins and neural development: a matter of targets and complexes. *Neuroreport* 2004, **15:**2567-2570.

14. Perrone-Bizzozero N, Bolognani F: Role of HuD and other RNA-binding proteins in neural development and plasticity. *J Neurosci Res* 2002, **68:**121-126.
15. Ule J, Jensen KB, Ruggiu M, Mele A, Ule A, Darnell RB: CLIP identifies Nova-regulated RNA networks in the brain. *Science* 2003, **302:**1212-1215.
16. Jensen KB, Dredge BK, Stefani G, Zhong R, Buckanovich RJ, Okano HJ, Yang YY, Darnell RB: Nova-1 regulates neuron-specific

alternative splicing and is essential for neuronal viability. *Neuron* 2000, **25:**359-371.
17. Larocque D, Galarneau A, Liu HN, Scott M, Almazan G, Richard S: Protection of p27(Kip1) mRNA by quaking RNA binding proteins promotes oligodendrocyte differentiation. *Nat Neurosci* 2005, **8:**27-33.
18. Sakakibara S, Nakamura Y, Yoshida T, Shibata S, Koike M, Takano H, Ueda S, Uchiyama Y, Noda T, Okano H: RNA-binding protein Musashi family: roles for CNS stem cells and a subpopulation of ependymal cells revealed by targeted disruption and antisense ablation. *Proc Natl Acad Sci U S A* 2002, **99:**15194-15199.
19. Pascale A, Gusev PA, Amadio M, Dottorini T, Govoni S, Alkon DL, Quattrone A: Increase of the RNA-binding protein HuD and posttranscriptional up-regulation of the GAP-43 gene during spatial memory. *Proc Natl Acad Sci U S A* 2004, **101:**1217-1222.
20. Jin P, Alisch RS, Warren ST: RNA and microRNAs in fragile X mental retardation. *Nat Cell Biol* 2004, **6:**1048-1053.
21. Dubnau J, Chiang AS, Grady L, Barditch J, Gossweiler S, McNeil J, Smith P, Buldoc F, Scott R, Certa U, Broger C, Tully T: The staufen/ pumilio pathway is involved in Drosophila long-term memory. *Curr Biol* 2003, **13:**286-296.
22. Miller S, Yasuda M, Coats JK, Jones Y, Martone ME, Mayford M: Disruption of dendritic translation of CaMKIIalpha impairs stabilization of synaptic plasticity and memory consolidation. *Neuron* 2002, **36:**507-519.
23. Kang H, Schuman EM: A requirement for local protein synthesis in neurotrophin-induced hippocampal synaptic plasticity. *Science* 1996, **273:**1402-1406.
24. Reymond A, Marigo V, Yaylaoglu MB, Leoni A, Ucla C, Scamuffa N, Caccioppoli C, Dermitzakis ET, Lyle R, Banfi S, Eichele G, Antonarakis SE, Ballabio A: Human chromosome 21 gene expression atlas in the mouse. *Nature* 2002, **420:**582-586.
25. Gray PA, Fu H, Luo P, Zhao Q, Yu J, Ferrari A, Tenzen T, Yuk DI, Tsung EF, Cai Z, Alberta JA, Cheng LP, Liu Y, Stenman JM, Valerius MT, Billings N, Kim HA, Greenberg ME, McMahon AP, Rowitch DH, Stiles CD, Ma Q: Mouse brain organization revealed through direct genome-scale TF expression analysis. *Science* 2004, **306:**2255-2257.
26. Gitton Y, Dahmane N, Baik S, Ruiz i Altaba A, Neidhardt L, Scholze M, Herrmann BG, Kahlem P, Benkahla A, Schrinner S, Yildirimman R, Herwig R, Lehrach H, Yaspo ML: A gene expression map of human chromosome 21 orthologues in the mouse. *Nature* 2002, **420:**586-590.
27. Saunders LR, Barber GN: The dsRNA binding protein family: critical roles, diverse cellular functions. *Faseb J* 2003, **17:**961-983.
28. Nagai K: RNA-protein complexes. *Curr Opin Struct Biol* 1996, **6:**53-61.
29. Burd CG, Dreyfuss G: Conserved structures and diversity of functions of RNA-binding proteins. *Science* 1994, **265:**615-621.
30. Wheeler DL, Church DM, Edgar R, Federhen S, Helmberg W, Madden TL, Pontius JU, Schuler GD, Schriml LM, Sequeira E, Suzek TO, Tatusova TA, Wagner L: Database resources of the National Center for Biotechnology Information: update. *Nucleic Acids Res* 2004, **32:**D35-40.
31. Sonnhammer EL, Eddy SR, Birney E, Bateman A, Durbin R: Pfam: multiple sequence alignments and HMM-profiles of protein domains. *Nucleic Acids Res* 1998, **26:**320-322.
32. Thomas PD, Campbell MJ, Kejariwal A, Mi H, Karlak B, Daverman R, Diemer K, Muruganujan A, Narechania A: PANTHER: a library of protein families and subfamilies indexed by function. *Genome Res* 2003, **13:**2129-2141.
33. **http://mahoney.chip.org/mahoney/RBP.** .
34. Buckanovich RJ, Posner JB, Darnell RB: Nova, the paraneoplastic Ri antigen, is homologous to an RNA-binding protein and is specifically expressed in the developing motor system. *Neuron* 1993, **11:**657-672.
35. Okano HJ, Darnell RB: A hierarchy of Hu RNA binding proteins in developing and adult neurons. *J Neurosci* 1997, **17:**3024-3037.
36. Kiehl TR, Shibata H, Vo T, Huynh DP, Pulst SM: Identification and expression of a mouse ortholog of A2BP1. *Mamm Genome* 2001, **12:**595-601.
37. Sakakibara S, Okano H: Expression of neural RNA-binding proteins in the postnatal CNS: implications of their roles in neuronal and glial cell development. *J Neurosci* 1997, **17:**8300-8312.

38. Sakakibara S, Nakamura Y, Satoh H, Okano H: Rna-binding protein Musashi2: developmentally regulated expression in neural precursor cells and subpopulations of neurons in mammalian CNS. *J Neurosci* 2001, **21:**8091-8107.
39. Blackshaw S, Harpavat S, Trimarchi J, Cai L, Huang H, Kuo WP, Weber G, Lee K, Fraioli RE, Cho SH, Yung R, Asch E, Ohno-Machado L, Wong WH, Cepko CL: Genomic analysis of mouse retinal development. *PLoS Biol* 2004, **2:**E247.
40. Lillevali K, Kulla A, Ord T: Comparative expression analysis of the genes encoding polypyrimidine tract binding protein (PTB) and its neural homologue (brPTB) in prenatal and postnatal mouse brain. *Mech Dev* 2001, **101:**217-220.
41. Zhang W, Liu H, Han K, Grabowski PJ: Region-specific alternative splicing in the nervous system: implications for regulation by the RNA-binding protein NAPOR. *Rna* 2002, **8:**671-685.
42. Grabowski PJ, Black DL: Alternative RNA splicing in the nervous system. *Prog Neurobiol* 2001, **65:**289-308.
43. Niehrs C, Pollet N: Synexpression groups in eukaryotes. *Nature* 1999, **402:**483-487.
44. Patton JG, Porro EB, Galceran J, Tempst P, Nadal-Ginard B: Cloning and characterization of PSF, a novel pre-mRNA splicing factor. *Genes Dev* 1993, **7:**393-406.
45. Hahm B, Cho OH, Kim JE, Kim YK, Kim JH, Oh YL, Jang SK:

Polypyrimidine tract-binding protein interacts with HnRNP L. *FEBS Lett* 1998, **425:**401-406.
46. Eperon IC, Makarova OV, Mayeda A, Munroe SH, Caceres JF, Hay-ward DG, Krainer AR: Selection of alternative 5' splice sites: role of U1 snRNP and models for the antagonistic effects of SF2/ASF and hnRNP A1. *Mol Cell Biol* 2000, **20:**8303-8318.
47. Hieronymus H, Silver PA: A systems view of mRNP biology. *Genes Dev* 2004, **18:**2845-2860.
48. Mody M, Cao Y, Cui Z, Tay KY, Shyong A, Shimizu E, Pham K, Schultz P, Welsh D, Tsien JZ: Genome-wide gene expression profiles of the developing mouse hippocampus. *Proc Natl Acad Sci U S A* 2001, **98:**8862-8867.
49. Mills JC, Andersson N, Hong CV, Stappenbeck TS, Gordon JI: Molecular characterization of mouse gastric epithelial progenitor cells. *Proc Natl Acad Sci U S A* 2002, **99:**14819-14824.
50. Stappenbeck TS, Hooper LV, Gordon JI: Developmental regulation of intestinal angiogenesis by indigenous microbes via Paneth cells. *Proc Natl Acad Sci U S A* 2002, **99:**15451-15455.
51. Harris MN, Ozpolat B, Abdi F, Gu S, Legler A, Mawuenyega KG, Tirado-Gomez M, Lopez-Berestein G, Chen X: Comparative proteomic analysis of all-trans-retinoic acid treatment reveals systematic posttranscriptional control mechanisms in acute promyelocytic leukemia. *Blood* 2004, **104:**1314-1323.

El fichero aquí mostrado es una conversión a fichero doc realizada por el Acrobat, durante la conversión se han perdido ciertas cosas como los pies de pagina y los encabezados, lo único que se pretende con la inserccion del texto es que sirva de muestra a la hora de comparar con las dos conversiones realizadas, dichas conversiones se realizaran manteniendo la distribución original del texto, por motivos evidentes.

El primero de los textos que se mostrara en texto plano será la conversión realizada por el XPDF y el segundo de los textos el APDF, luego se escribirán unas breves conclusiones y asi se dara por concluido este anexo.

Conversion con el XPDF

BMC Developmental Biology

Research article

BioMed Central

Open Access

A genome-wide in situ hybridization map of RNA-binding proteins reveals anatomically restricted expression in the developing mouse brain

Adrienne E McKee1,2, Emmanuel Minet2,3, Charlene Stern2, Shervin Riahi2, Charles D Stiles2,4 and Pamela A Silver*1,2

Address: 1Department of Systems Biology, Harvard Medical School, Boston, MA 02115 USA, 2Department of Cancer Biology, The Dana-Farber Cancer Institute, Boston, MA 02115 USA, 3URBC-FUNDP, 61 rue de Bruxelles, 5000 Namur, Belgium and 4Department of Microbiology and Molecular Genetics, Harvard Medical School, Boston, MA 02115 USA Email: Adrienne E McKee - adrienne_mckee@student.hms.harvard.edu; Emmanuel Minet - emmanuel.minet@fundp.ac.be; Charlene Stern - cstern@foleyhoag.com; Shervin Riahi - shervin@gmail.com; Charles D Stiles - charles_stiles@dfci.harvard.edu; Pamela A Silver* - pamela_silver@dfci.harvard.edu * Corresponding author Equal contributors

Published: 20 July 2005 BMC Developmental Biology 2005, 5:14 doi:10.1186/1471-213X-5-14

Received: 06 May 2005 Accepted: 20 July 2005

Abstract

Background: In eukaryotic cells, RNA-binding proteins (RBPs) contribute to gene expression by regulating the form, abundance, and stability of both coding and non-coding RNA. In the vertebrate brain, RBPs account for many distinctive features of RNA processing such as activity-dependent transcript localization and localized protein synthesis. Several RBPs with activities that are important for the proper function of adult brain have been identified, but how many RBPs exist and where these genes are expressed in the developing brain is uncharacterized. Results: Here we describe a comprehensive catalogue of the unique RBPs encoded in the mouse genome and provide an online database of RBP expression in developing brain. We identified 380 putative RBPs in the mouse genome. Using in situ hybridization, we visualized the expression of 323 of these RBP genes in the brains of developing mice at embryonic day 13.5, when critical fate choice decisions are made and at P0, when major structural components of the adult brain are apparent. We demonstrate i) that 16 of the 323 RBPs examined show neural-specific expression at the stages we examined, and ii) that a far larger subset (221) shows regionally restricted expression in the brain. Of the regionally restricted RBPs, we describe one group that is preferentially expressed in the E13.5 ventricular areas and a second group that shows spatially restricted expression in postmitotic regions of the embryonic brain. Additionally, we find a subset of RBPs that share the same complex pattern of expression, in proliferating regions of the embryonic and postnatal NS and peripheral tissues. Conclusion: Our data show that, in contrast to their proposed ubiquitous involvement in gene regulation, most RBPs are not uniformly expressed. Here we demonstrate the region-specific expression of RBPs in proliferating vs. post-mitotic brain regions as well as cell-type-specific RBP expression. We identify uncharacterized RBPs that exhibit neural-specific expression as well as novel RBPs that show expression in non-neural tissues. The data presented here and in an online database provide a visual filter for the functional analysis of individual RBPs.

Background

The ordered production and differentiation of cell types that occurs during nervous system (NS) development relies upon tightly regulated gene expression. In neural cells, spatial and temporal gene regulation occurs through both transcriptional and post-transcriptional mechanisms. While the transcriptional networks that direct neural cell fate and govern cell shape, position, and connectivity have been well studied [1-3], the post-transcriptional influences on neural development and gene expression are less well understood. At the core of post-transcriptional gene regulation are RNA-binding proteins (RBPs). Proteins containing canonical RNA-binding domains (RBDs) are involved in numerous steps of nuclear and cytoplasmic RNA processing [4]. Through mRNA capping, splicing, editing, polyadenylation and nonsense-mediated decay, RBPs modulate the diversity of transcribed genes [4-6]. RBPs also affect the processing of non-coding RNAs [7]. Specific RBPs additionally enable asymmetric RNA distribution and translational regulation [8-10], two phenomena that are critical for affecting localized protein synthesis [11,12]. The importance of post-transcriptional processing in NS gene regulation is underscored by functional examples of specific RBPs [13,14]. For instance, the neuronal-specific factor Nova-1 regulates splicing of pre-mRNAs that encode components of inhibitory synapses [15]. Mice lacking Nova-1 die postnatally due to aberrant regulation of apoptotic neuronal death [16]. As a second example, RBPs encoded by the quaking and Musashi loci promote glial cell fate [17] and CNS stem cell self-renewal [18] by stabilizing transcripts involved in cell differentiation. Thirdly, the fragile X mental retardation protein, members of the ELAV/Hu protein family, and the Staufen proteins are involved in targeting and translational regulation of dendritic transcripts [19-21]. Additionally, the finding that long-term memory requires de novo protein synthesis highlights the significance of post-transcriptional processes in neural function [22,23]. Despite our knowledge of several key RBPs, much of the understanding of RBPs in the brain comes from studies of adult animals or neural cell lines. Thus, how the functional class of RBPs contributes to the positioning, growth, and diversification of cells in the developing brain is not well understood. One step towards increasing our understanding RBPs is to resolve where they are expressed. Here, we utilize the approach of in situ hybridization mapping [24-26] to investigate the expression of 323 RBPs within the developing mouse brain. Two stages of development were characterized, embryonic day 13.5 (E13.5), when critical cellular fate choice decisions are made and postnatal day 0 (P0), when the major structural

components of the brain are apparent. We find that, in contrast to their proposed ubiquitous involvement in gene regulation, most RBPs are not uniformly expressed. The majority of RBPs profiled demonstrates spatially restricted expression in the brain or in other peripheral tissues examined. The data presented here and in an online database afford a visual filter for the functional analysis of individual RBPs in the developing mammalian NS.

Results

Mouse RBPs were identified according to gene sequence The RNA recognition motif (RRM), the hnRNP K-homology (KH) domain, and the double-stranded RNA-binding domain (dsRM) are evolutionarily conserved, well-characterized domains known to bind either single or doublestranded RNA [27-29]. Sequence similarity searches and structural analyses of these domains have led to the ability to predict other RBPs based on primary coding sequence [29]. To identify unique genomic loci that encode putative RBPs in the mouse genome, we analyzed existing public [30,31] and private [32] databases for sequences containing one or more RBD. Candidates were classified as RBPs only if their predicted protein sequence contained a Protein Families Database (Pfam)-defined RBD [31].

We identified 290 genes harboring one or more RRM, KH, or dsRM sequences. We also identified 32 genes encoding other domains shown to interact with RNA, including the zinc knuckle, G-patch, PIWI, DEAD box RNA helicase, and TUDOR domains. Finally, as the absence of a canonical RBD does not preclude interaction with RNA, we sought 58 additional genes known or predicted to be associated with RNA processing. In total, this collection contains 380 putative RBPs. Additional file 1 lists the number of genes, per RBD, identified and analyzed by in situ hybridization. A list of all genes and primer sequences is given in Additional file 2.

RBP expression in the developing mouse brain was analyzed by in situ hybridization To localize RBP expression, we preformed in situ hybridization on whole head tissue sections of E13.5 embryos and P0 mice. We designed gene-specific primers to produce 400700 bp probes for 340 candidate RBPs. These primer sets were used to perform PCR on cDNA prepared from embryonic or P0 mouse brains. A small number of probes were obtained from mouse intestine, liver, kidney, or testes cDNA. 323 genes (95%) showed positive PCR products (data not shown). Following subcloning, antisense digoxygenin-labeled riboprobes were prepared and hybridized against coronal head and transverse upperbody sections (to include the brain and spinal cord, respectively). Digital images of the entire in situ hybridization set have been deposited in the Mahoney RNA-Binding Protein Expression Database [33].

Figure 1 RBP expression in proliferative zones of the E13.5 mouse forebrain RBP expression in proliferative zones of the E13.5 mouse forebrain. In situ hybridization patterns for four RBPs on sections through the forebrain of E13.5 mice. Labels indicate Locuslink gene names. All images show the same magnification.

RBPs exhibit restricted expression in the developing mouse brain Several neural-specific RBPs have been identified, yet how many others demonstrate this degree of specificity is unknown. Of the genes examined we found 16 RBPs (listed in Additional file 2) that exhibit NS-restricted expression in the tissues analyzed. Among this list are known examples of neuronal-specific RBPs including Nova-1 [34], the ELAV/Hu proteins B, C, and D [35], and Ataxin 2 binding protein 1 (A2bp1) [36] but additionally include putative RBPs for which expression has not been reported. With the exception of one gene that was only detected at E13.5, all (15/16) of these RBPs appear brain or NS-specific at both developmental stages in the tissues analyzed. Overall, these RBP encoding genes are not limited in expression to one brain region but are found in multiple brain or NS structures. RBPs show spatially restricted expression in anatomically distinct brain regions We find that greater than half of the RBPs profiled exhibit spatially restricted expression. Of the 323 genes examined, 221 demonstrate localized, enriched expression in one or more discrete brain regions in addition to detectable expression in non-NS tissues. We divided the E13.5 and P0 CNS into five and eight general areas for annotation, respectively: the E13.5 precortical area, the striatum (and other basal ganglia), the periventrical areas, hindbrain, and spinal cord, as well as the P0 cortex, striatum, hippocampus, thalamus, hypothalamus, midbrain, hindbrain, and spinal cord. The presence or absence of expression for each RBP was analyzed visually at each location and is annotated in Additional file 3. Very few of the 221

RBPs with spatially restricted expression patterns were expressed in only one brain region, however most (73%) showed restricted expression at both developmental stages (Additional file 3). We observe multiple RBPs that demonstrate region-specific expression in the E13.5 ventricular areas. Shown in Figure 1 are representative RBP genes that are transcribed in mitotically-active cells in the neuroepithelia of the developing telencephalon. Among the RBPs expressed in this region occupied by neural progenitor cells, we find examples of mRNA export factors in addition to putative splicing factors and transcriptional regulators (Fig. 1). In all instances, expression in the embryonic lateral ventricular zone is accompanied by expression in the periventricular areas of the 3rd and 4th E13.5 ventricles and often by heightened expression in the P0 subventricular zone [33]. Notably, we observed this pattern of expression for the dsRM-containing Musashi proteins [33]. Our results are consistent with the documented expression of Msi1 and Msi2 [37,38]. Multiple RBPs show restricted expression in post-mitotic regions of embryonic brain. Presented in Figure 2 are examples of four putative RBPs that demonstrate regionspecific expression in areas containing post-mitotic neurons. Transcripts of the genes encoding the RRM protein Brunol6 and the predicted zinc-knuckle protein 1500031H04Rik appear pan-neuronal at both developmental stages (Fig. 2A, 2B and [33]). Expression of the RRM-containing RIKEN gene 4930565A21 is most pronounced in the ventral telencephalon, while D11Bwg0517e is found in the precortical layer, the

Figure 2 RBP expression in post-mitotic areas of the E13.5 mouse forebrain RBP expression in post-mitotic areas of the E13.5 mouse forebrain. In situ hybridization patterns for four RBPs on sections through the forebrain of E13.5 mice. Labels indicate Locuslink gene names. bg, basal ganglia; hy, hypothalamus; nc, neocortex. All images show the same magnification.

thalamic area and hindbrain (Fig. 2C, 2D and [33]). Among the genes that occupy post-mitotic regions of the developing brain we additionally observe members of the ELAV/Hu family as well as other RBPs that have well-documented neuronal expression [34,35].

RBPs demonstrate cell-type specific expression in the P0 mouse retina As our in situ hybridization analyses were performed on sections through whole head, we were able to visualize RBP expression in the developing retina. The vertebrate retina provides a distinctive system for studying CNS development as its seven major neural cell types are readily distinguished from one another by their morphology and laminar position [39]. Shown in Figure 3 are examples of the diversity of RBP expression in the P0 retina. The RRM-containing A2bp1 is expressed in the retinal ganglion cell layer (GCL), which contains primarily retinal ganglion cells and a small number of displaced amacrine cells (Fig 3A, 3B). The KH-domain encoding gene poly(rC) binding protein 3 (Pcbp3) shows dramatically enriched expression in the inner nuclear layer (INL) (Fig. 3C and 3D), possibly indicating localization to the bipolar neuron cell bodies that occupy the scleral portion of the INL. Notably, both A2bp1 and Pcbp3 show restricted expression in post-mitotic regions of the E13.5 and P0 brain [24,36]. Transcripts of the RRM-encoding scaffold attachment factor B (Safb) and of the three-RRM containing SPOC gene Rbm15 are expressed in the outer neuroblastic layer of the retina (Fig. 3EH). Safb, but not Rbm15, is additionally expressed in the GCL, possibly in the Müller glia. Both Safb and Rbm15 show enriched expres-

sion in neuroepithelia of the ventricular zone (Fig. 1 and [33]).

A systems-based view of RBP expression Gene regulation by RBPs is believed to occur through coordinated, combinatorial interactions with RNA. During the course of this study we identified multiple RBPs that are coordinately expressed in the brain and other tissues. We find 48 genes (listed in Additional file 4) that show elevated expression in proliferating areas of the embryonic and postnatal brain as well as in postnatal nasal epithelia, teeth, and thymus. Presented in Figure 4 are expression data for snRNP E and Son, two representative examples of this "synexpression group" of genes that share a similar, complex pattern of expression. Further examples are shown in Additional file 5. This same expression distribution has been observed for the polypyrimidine tract-binding protein, PTBP1, and our data are consistent with previous findings [40]. Notably, the protein products of many of the genes listed are understood to interact either physically or genetically. RBPs show restricted expression in non-NS tissues As our analyses were performed on whole head and upper thoracic tissues, our data provide detailed information about RBP expression in developing cranial facial tissues. We identified putative RBPs that display tissue-restricted expression in non-NS structures (listed in Additional file 3). Figure 5 presents in situ hybridization results for two RRM-encoding transcripts that show highly restricted expression in different epithelial tissues. The Riken gene 2210008M09 is transcribed in epithelia covering the facial skeleton (Fig. 5A, 5B), while the gene BC013481 is

Figure 3 P0 retina Diversity of RBP expression in major cellular subtypes of the Diversity of RBP expression in major cellular subtypes of the P0 retina. In situ hybridization for four representative RBPs that exhibit laminar-specific expression in the P0 mouse retina. Labels indicate Locuslink gene names. A, B) A2bp1, C, D) Pcbp3, E, F) Safb, G, H) Rbm15. Panels A, C, E, and G show the same magnification. Panels B, D, F, and H show the same magnification. gcl, granule cell layer; inl, inner nuclear layer, onbl; outer neuroblastic layer.

expressed in the choroid plexus (Fig. 5C) and in the lining of the intestine and placenta (Fig. 5D, 5E).

Discussion

Neural cells utilize multiple forms of post-transcriptional gene regulation. While RBPs are believed to be potent modulators of post-transcriptional processes, little is known about how this functional class is expressed in the developing brain. As a first step towards increasing our knowledge of RBPs we chose to investigate the spatial and temporal expression of genes that encode motifs known

Figure 4 P0 mouse tissues Representative examples of RBP synexpression in E13.5 and Representative examples of RBP synexpression in E13.5 and P0 mouse tissues. snRNP E and Son are transcribed in the perventricular areas of the E13.5 brain (A, E), in the P0 subventricular area of the lateral ventricle (B, F), in the external granule layer of the P0 cerebellum (C, G), as well as in postnatal developing teeth (D, H).

to interact with RNA. We find a small set of RBPs that show neural-specific expression in the tissues analyzed.

brain (Fig. 2). Based on their pattern of expression, these RBPs may have roles in neural proliferation, cell fate choice and cell migration, or in neuronal function, respectively. We also identified novel RBPs that are expressed in tissues of mesodermal and endodermal origin (Fig. 5). The highly restricted expression of these genes may indicate an explicit role for these RBPs in their respective epithelia. Additionally, the cell-type specificity RBPs found in the P0 retina (Fig. 3) illustrates the diversity of RBP expression. The specialized expression of these RBPs may be indicative of a dedicated function in the specified tissues. By visual inspection of in situ hybridization data, we find a subset of RBPs that are coordinately expressed in multiple tissue types. These genes display heightened expression in the periventricular areas of the E13.5 brain and spinal cord as well as marked expression in the external granule layer of the P0 cerebellum, the lateral subventricular zones, and in teeth, nasal epithelia, and thymus (Fig. 4, Additional file 5, [33]). While not excluded from postmitotic tissues, these RBPs are predominately expressed in structures that are undergoing cell division. Notably, the term 'synexpression group' has been used to describe collections of genes that function in a common process and share a similar complex spatial expression pattern in multiple tissues [43]. Among the synexpression group identified here we find examples of RBPs that are known to interact either physically or genetically (Additional file 4). For example, PTBP1 binds the splicing factors PSF [44] and hnRNP L [45] while SF2/ASF and hnRNP A1 select for 5' exon or exclusion or inclusion, respectively [46]. Our data provide visual support to a growing body of evidence that functionally-related transcripts are post-transcriptionally co-regulated [47]. Although the significance of certain splicing and mRNA export factor enrichment in proliferating regions is not known, data from multiple studies point to a role for RBPs in cell proliferation. During hippocampal development expression levels of RBPs were found to be high and then to dramatically decrease, as neurons transition from a proliferating to a post-mitotic state [48]. A number of RBPs were also identified as highly expressed in a molecular characterization of gastric epithelial progenitor cells [49,50]. Furthermore, protein levels of hnRNPs and snRNPs were found to be down-regulated upon stimulated growth inhibition of myeloid cells [51]. Therefore, it is likely that a role for RBPs during cell proliferation and cell fate determination exists in multiple tissue types.

Figure 5 restricted expression of novel RBPs the non-neural, In situ hybridization profiling uncovers In situ hybridization profiling uncovers the non-neural, restricted expression of novel RBPs. Data from ISH performed on (A, C) coronal E13.5 and on (B, D, E) E15 sagittal sections are presented for RRM-encoding RBPS. A, B) The Riken gene 2210008M09 is transcribed in epithelia covering the facial skeleton. C-E) BC013481 is detected in the choroid plexus, in the intestinal lining, and in the lining of the placenta. Panels C-E show the same magnification.

An even greater number of RBP genes however demonstrate spatially restricted expression in distinct regions of the developing brain. Within the CNS, most of the RBPs examined show nonuniform, heightened expression in anatomically discrete structures. Tissue differences in the expression levels of individual genes could indicate distinctive protein requirements among cell types, beyond that of tissue-specific RBPs [41]. There is precedent for differential requirements of individual RBPs, as tissue-specific RNA splicing is achieved partly through combinatorial, stoichiometric differences among splicing factors within various cells [42]. It is from this local enrichment within different cell types or tissues that we can begin to hypothesize as to the functional significance of individual genes as well as to the importance of groups of similarly expressed RBPs. Our study has identified RBPs that display spatially restricted expression in distinct regions of the developing mouse brain. One set of RBPs (Fig. 1) is found in the E13.5 ventricular areas. A second set demonstrates spatially restricted expression in post-mitotic regions of E13.5

Conclusion

In summary, the data presented here provide new insight into how a distinct functional gene class is expressed in the developing NS. We find that RBPs demonstrate

region-specific as well as cell-type specific expression. In addition, we find that specific, proliferating regions of the embryonic and postnatal NS and peripheral tissues are similar in the expression of certain RBPs. These data serve as a starting point for functional investigations into the roles of RBPs in neural development and physiology.

Methods

In silico RBP identification Putative RBP gene sequences were identified by homology-based whole genome screening using public and private databases: Celera Panther Families, Protein Families Database (Pfam), and Genbank [30-32]. Classification as an RBP was based on the presence of one or more RRM, KH, or dsRMs, as defined by Pfam databases [31]. Databases were also mined for zinc-knuckle, G-patch, PIWI, DEAD-box helicase and Tudor domain-containing sequences and for known factors involved in mRNA splicing, editing, transport, and stability. Genes with multiple RNA-binding domains were assigned to a single subfamily. Unique gene identity was verified by LocusID numbers. As of March 1, 2004, a total of 357 unique genes were identified from these sources. An additional 26 RRM, KH, and dsRM proteins have been identified as of March 7, 2005. PCR primer design PCR primer pairs were designed for each identified RNAbinding protein locus. PCR primer sequences were designed with approximately 60% GC content, spanning 400700 base pairs of primarily the gene's coding sequence. Additional primer pairs were designed for targets that did not initially yield PCR products. Cloning Total RNA was obtained from E13.5, P0, or adult C57/BL6 mouse brains (Charles River Laboratories) by Trizol extraction (Invitrogen). Reverse transcription was performed using Superscript II reverse transcriptase and oligo-dT (Invitrogen). PCR was performed with cDNA templates using 40 cycles, 6065°C annealing temperature, and Platinum Taq (Invitrogen) as polymerase. For a few genes, PCR was performed with cDNA templates prepared from adult brain, kidney, gut, liver, or testis tissues. Positive PCR products were cloned into TA cloning vectors (Invitrogen) and verified by restriction digest or DNA sequencing. Probe synthesis Gene fragments from verified plasmids were amplified by PCR using plasmid specific primers. Digoxigenin-labeled RNA probes were made, using PCR products as template and T7 or SP6 RNA polymerases (Roche). cRNA probes were ethanol precipitated and quantified by spectrophotometry.

Tissue preparation E13.5 embryos were directly fixed overnight in 4% paraformaldehyde (0.1M PBS). P0 mice were transcardially perfused with 4% paraformaldehyde (0.1M PBS) and postfixed overnight at 4°C. After fixation, embryos and P0 mice were transferred to 20% sucrose overnight. The head, neck, and trunk were embedded separately in OCT (Tissue-Tek) on dry ice and stored at -80°C. Serial cryostat sections (14 μm) were cut and mounted on Superfrost Plus slides (Fisher). Ten and twenty adjacent sets of sections were prepared from E13.5 embryos and P0 mice, respectively, and were stored at -20°C until use. Section in situ hybridization In situ hybridization was performed according to Gray et al. [25]. Following pretreatment (Proteinase K), slides were pre-hybridized for 1h at 65°C in hybridization solution (50% formamide (Ambion), 5X SSC, 0.3 mg/ml yeast tRNA (Sigma), 100 μg/ml heparin (Sigma), 1X Denhardt's (Sigma), 0.1% tween, 5 mM EDTA). P0 and E13.5 brain sections were hybridized overnight with labeled RNA probe(0.81.2 μg/ml) at 65°C, washed in 2X SSC at 67°C, incubated with RNase A (1 μg/ml, 2X SSC) at 37°C, washed in 0.2X SSC at 65°C, blocked in PBS with 10% lamb sera, and incubated in alkaline phosphatase labeled anti-DIG antibody (Roche) (1:2000, 10% sera) overnight. Sections were washed and color was visualized using NBT and BCIP in alkaline phosphatase buffer (100 mM Tris pH 9.5, 50 mM MgCl2, 100 mM NaCl, 0.1% tween-20) containing 75 μg/ml NBT (BioRad), 600 μg/ml BCIP (Roche). Staining was stopped after visual inspection. Sections were washed, fixed in 4% paraformaldehyde, and coverslipped in glycerol [25]. Image acquisition and RBP expression database Images were acquired and analyzed as described [25]. Images were either scanned using a Nikon Coolscan 8000 slide scanner (4000 DPI) or digitally acquired using a Leica digital camera. Image levels have been modified in Photoshop (Adobe) for clarity. Full resolution scanned images were compressed using JPEG compression, quality 10, and have been deposited in the Mahoney RNA-Binding Protein Expression Database [33].

Authors' contributions

AEM prepared tissue samples, performed data analysis and drafted the manuscript. EM performed data analysis and both EM and SR generated reagents, tissue samples, digitized the raw data, and helped build the website. CS contributed to the design of the study and prepared tissue samples. CDS and PAS conceived of the study, participated in its design and coordination and helped prepare the manuscript. All authors read and approved of the manuscript.

BMC Developmental Biology 2005, 5:14

http://www.biomedcentral.com/1471-213X/5/14

Additional material Additional File 1

RNA-binding proteins identified in silico and profiled by in situ hybridization. List of annotated RNA-binding domains and the number of family members that were identified in silico and analyzed by in situ hybridization. Click here for file [http://www.biomedcentral.com/content/supplementary/1471213X-5-14-S1.xls]

Additional File 5

Examples of RBP synexpression in E13.5 and P0 mouse tissues. Additional examples of RBPs that share a similar pattern of expression. Shown are in situ hybridization results of expression in the periventricular areas of the E13.5 brain (A, E, I, M, Q), in the subventricular area of the P0 lateral ventricle (B, F, J, N, R), in the external granule layer of the P0 cerebellum (C, G, K, O, S), as well as in postnatal developing teeth (D, H, L P, T). A-D) Refbp1, E-H) hnRNP A1, I-L) PTBP1, M-P) Sfpq, QR) Hnrpl. Panels A, B, E, F, I, J, M, N, Q, R show the same magnification. Panels C, D, G, H, K, L, O, P, S, T show the same magnification. Click here for file [http://www.biomedcentral.com/content/supplementary/1471213X-5-14-S5.png]

Additional File 2

List of 380 genes identified as putative RBPs in the mouse genome and analyzed in this study. Columns indicate LocusID, gene name, type of RBD, primer sequences used to isolate the target cDNA, the size of the cDNA fragment, the presence call by PCR from E13.5 and P0 brain cDNA, cloning status ('c' indicates cloned, 'u' indicates uncloned, 'small' indicates that the target gene had less than 400 bp of unique sequence, 'na' indicates that cloning was not attempted), the RNA polymerase used to generate the anti-sense riboprobe, the tissue from which the cDNA was isolated (if not from E13.5 or P0 mouse brain), and whether the gene was analyzed by in situ hybridization ('x' indicates yes). Click here for file [http://www.biomedcentral.com/content/supplementary/1471213X-5-14-S2.xls]

Acknowledgements

We are grateful to Drs. Qiufu Ma and John Alberta for critical review of this manuscript and for assistance in this work. We thank Eric Tsung, Zhaohui Cai, and Matthew McCormack for designing the website. This work has been supported by the Bernard A. and Wendy J. Goldhirsh Foundation for Brain Tumor Research and by the Charles A. Dana Foundation. AEM is supported by an institutional training grant from the National Cancer Institute (T32CA09361). EM is funded as a FNRS Researcher through the Belgian National Research Fund and by the D. Collen Research Foundation VZW and BAEF. CS received support from the American Cancer Society (PF-02128-01-MBC).

Additional File 3

Complete list of gene expression patterns for all in situ hybridizations performed. Of the 323 RBPs examined, 221 showed restricted expression patterns in the brain. The remaining genes either show restricted expression in non-neural tissues, ubiquitous expression that is difficult to distinguish from background, or no expression. Caution is needed in interpreting the results. First, non-expression could be due to the sensitivity limit of non-radioactive in situ hybridization. Second, the background level of individual probes may differ. Third, some probes with high background hybridization may mask the real expression of the transcript. Fourth, we cannot rule out the possibility that some probes may show variable levels of background hybridization in different brain areas, resulting in a false positive signal. Columns A-D describe the LocusID, gene name, type of RBD, and number (internal Mahoney reference number). Columns E and, L (E13.5, P0 "Informativity"): "1" for restricted expression in the nervous system and "0" for either ubiquitous expression that is difficult to distinguish from background or no expression. As noted in Gray et al [25], some of the genes in the "0" category show uneven signals in different brain regions and are also annotated in the subsequent columns. Columns F and M (E13.5, P0 "Specificity"): "1" for restricted expression in neural tissues only, "2" for restricted expression in neural tissue with distinguishable expression in non-neural tissue, "3" for ubiquitous or no expression, and "4" for expression in non-neural tissues only. Columns GK and N-U (E13.5, P0 "Expression"): "2" for expression, "1" for ubiquitous expression or background, "0" for no expression. Click here for file [http://www.biomedcentral.com/content/supplementary/1471213X-5-14-S3.xls]

References

1. 2. 3. 4. 5. 6. 7. 8. 9. 10. 11. 12. 13. 14. 15. 16. Ross SE, Greenberg ME, Stiles CD: Basic helix-loop-helix factors in cortical development. Neuron 2003, 39:13-25. Wilson SW, Houart C: Early steps in the development of the forebrain. Dev Cell 2004, 6:167-181. Bally-Cuif L, Hammerschmidt M: Induction and patterning of neuronal development, and its connection to cell cycle control. Curr Opin Neurobiol 2003, 13:16-25. Dreyfuss G, Kim VN, Kataoka N: Messenger-RNA-binding proteins and the messages they carry. Nat Rev Mol Cell Biol 2002, 3:195-205. Lasko P: Gene regulation at the RNA layer: RNA binding proteins in intercellular signaling networks. Sci STKE 2003, 2003:RE6. Orphanides G, Reinberg D: A unified theory of gene expression. Cell 2002, 108:439-451. Cullen BR: Transcription and processing of human microRNA precursors. Mol Cell 2004, 16:861-865. Huang YS, Carson JH, Barbarese E, Richter JD: Facilitation of dendritic mRNA transport by CPEB. Genes Dev 2003, 17:638-653. Antar LN, Bassell GJ: Sunrise at the synapse: the FMRP mRNP shaping the synaptic interface. Neuron 2003, 37:555-558. Tang SJ, Meulemans D, Vazquez L, Colaco N, Schuman E: A role for a rat homolog of staufen in the transport of RNA to

neuronal dendrites. Neuron 2001, 32:463-475. Martin KC: Local protein synthesis during axon guidance and synaptic plasticity. Curr Opin Neurobiol 2004, 14:305-310. Huang YS, Richter JD: Regulation of local mRNA translation. Curr Opin Cell Biol 2004, 16:308-313. Agnes F, Perron M: RNA-binding proteins and neural development: a matter of targets and complexes. Neuroreport 2004, 15:2567-2570. Perrone-Bizzozero N, Bolognani F: Role of HuD and other RNAbinding proteins in neural development and plasticity. J Neurosci Res 2002, 68:121-126. Ule J, Jensen KB, Ruggiu M, Mele A, Ule A, Darnell RB: CLIP identifies Nova-regulated RNA networks in the brain. Science 2003, 302:1212-1215. Jensen KB, Dredge BK, Stefani G, Zhong R, Buckanovich RJ, Okano HJ, Yang YY, Darnell RB: Nova-1 regulates neuron-specific

Additional File 4

RNA-binding proteins belonging to a synexpression group. Complete list of RBPs that demonstrate a similar complex pattern of expression. Columns A-D describe the LocusID, gene name, type of RBD, and number (internal Mahoney reference number). Click here for file
[http://www.biomedcentral.com/content/supplementary/1471213X-5-14-S4.xls]

17.

18.

19.

20. 21.

22.

23. 24.

25.

26.

27. 28. 29. 30.

31. 32.

33. 34.

35. 36. 37.

alternative splicing and is essential for neuronal viability. Neuron 2000, 25:359-371. Larocque D, Galarneau A, Liu HN, Scott M, Almazan G, Richard S: Protection of p27(Kip1) mRNA by quaking RNA binding proteins promotes oligodendrocyte differentiation. Nat Neurosci 2005, 8:27-33. Sakakibara S, Nakamura Y, Yoshida T, Shibata S, Koike M, Takano H, Ueda S, Uchiyama Y, Noda T, Okano H: RNA-binding protein Musashi family: roles for CNS stem cells and a subpopulation of ependymal cells revealed by targeted disruption and antisense ablation. Proc Natl Acad Sci U S A 2002, 99:15194-15199. Pascale A, Gusev PA, Amadio M, Dottorini T, Govoni S, Alkon DL, Quattrone A: Increase of the RNA-binding protein HuD and posttranscriptional up-regulation of the GAP-43 gene during spatial memory. Proc Natl Acad Sci U S A 2004, 101:1217-1222. Jin P, Alisch RS, Warren ST: RNA and microRNAs in fragile X mental retardation. Nat Cell Biol 2004, 6:1048-1053. Dubnau J, Chiang AS, Grady L, Barditch J, Gossweiler S, McNeil J, Smith P, Buldoc F, Scott R, Certa U, Broger C, Tully T: The staufen/ pumilio pathway is involved in Drosophila long-term memory. Curr Biol 2003, 13:286-296. Miller S, Yasuda M, Coats JK, Jones Y, Martone ME, Mayford M: Disruption of dendritic translation of CaMKIIalpha impairs stabilization of synaptic plasticity and memory consolidation. Neuron 2002, 36:507-519. Kang H, Schuman EM: A requirement for local protein synthesis in neurotrophin-induced hippocampal synaptic plasticity. Science 1996, 273:1402-1406. Reymond A, Marigo V, Yaylaoglu MB, Leoni A, Ucla C, Scamuffa N, Caccioppoli C, Dermitzakis ET, Lyle R, Banfi S, Eichele G, Antonarakis SE, Ballabio A: Human chromosome 21 gene expression atlas in the mouse. Nature 2002, 420:582-586. Gray PA, Fu H, Luo P, Zhao Q, Yu J, Ferrari A, Tenzen T, Yuk DI, Tsung EF, Cai Z, Alberta JA, Cheng LP, Liu Y, Stenman JM, Valerius MT, Billings N, Kim HA, Greenberg ME, McMahon AP, Rowitch DH, Stiles CD, Ma Q: Mouse brain organization revealed through direct genome-scale TF expression analysis. Science 2004, 306:2255-2257. Gitton Y, Dahmane N, Baik S, Ruiz i Altaba A, Neidhardt L, Scholze M, Herrmann BG, Kahlem P, Benkahla A, Schrinner S, Yildirimman R, Herwig R, Lehrach H, Yaspo ML: A gene expression map of human chromosome 21 orthologues in the mouse. Nature 2002, 420:586-590. Saunders LR, Barber GN: The dsRNA binding protein family: critical roles, diverse cellular functions. Faseb J 2003, 17:961-983. Nagai K: RNA-protein complexes. Curr Opin Struct Biol 1996, 6:53-61. Burd CG, Dreyfuss G: Conserved

structures and diversity of functions of RNA-binding proteins. Science 1994, 265:615-621. Wheeler DL, Church DM, Edgar R, Federhen S, Helmberg W, Madden TL, Pontius JU, Schuler GD, Schriml LM, Sequeira E, Suzek TO, Tatusova TA, Wagner L: Database resources of the National Center for Biotechnology Information: update. Nucleic Acids Res 2004, 32:D35-40. Sonnhammer EL, Eddy SR, Birney E, Bateman A, Durbin R: Pfam: multiple sequence alignments and HMM-profiles of protein domains. Nucleic Acids Res 1998, 26:320-322. Thomas PD, Campbell MJ, Kejariwal A, Mi H, Karlak B, Daverman R, Diemer K, Muruganujan A, Narechania A: PANTHER: a library of protein families and subfamilies indexed by function. Genome Res 2003, 13:2129-2141. http://mahoney.chip.org/mahoney/RBP. . Buckanovich RJ, Posner JB, Darnell RB: Nova, the paraneoplastic Ri antigen, is homologous to an RNA-binding protein and is specifically expressed in the developing motor system. Neuron 1993, 11:657-672. Okano HJ, Darnell RB: A hierarchy of Hu RNA binding proteins in developing and adult neurons. J Neurosci 1997, 17:3024-3037. Kiehl TR, Shibata H, Vo T, Huynh DP, Pulst SM: Identification and expression of a mouse ortholog of A2BP1. Mamm Genome 2001, 12:595-601. Sakakibara S, Okano H: Expression of neural RNA-binding proteins in the postnatal CNS: implications of their roles in neuronal and glial cell development. J Neurosci 1997, 17:8300-8312.

38.

39.

40.

41. 42. 43. 44. 45. 46.

47. 48.

49. 50. 51.

Sakakibara S, Nakamura Y, Satoh H, Okano H: Rna-binding protein Musashi2: developmentally regulated expression in neural precursor cells and subpopulations of neurons in mammalian CNS. J Neurosci 2001, 21:8091-8107. Blackshaw S, Harpavat S, Trimarchi J, Cai L, Huang H, Kuo WP, Weber G, Lee K, Fraioli RE, Cho SH, Yung R, Asch E, Ohno-Machado L, Wong WH, Cepko CL: Genomic analysis of mouse retinal development. PLoS Biol 2004, 2:E247. Lillevali K, Kulla A, Ord T: Comparative expression analysis of the genes encoding polypyrimidine tract binding protein (PTB) and its neural homologue (brPTB) in prenatal and postnatal mouse brain. Mech Dev 2001, 101:217-220. Zhang W, Liu H, Han K, Grabowski PJ: Region-specific alternative splicing in the nervous system: implications for regulation by the RNA-binding protein NAPOR. Rna 2002, 8:671-685. Grabowski PJ, Black DL: Alternative RNA splicing in the nervous system. Prog Neurobiol 2001, 65:289-308. Niehrs C, Pollet N: Synexpression groups in eukaryotes. Nature 1999, 402:483-487. Patton JG, Porro EB, Galceran J, Tempst P, Nadal-Ginard B: Cloning and characterization of PSF, a novel pre-mRNA splicing factor. Genes Dev 1993, 7:393-406. Hahm B, Cho OH, Kim JE, Kim YK, Kim JH, Oh YL, Jang SK: Polypyrimidine tract-binding protein interacts with HnRNP L. FEBS Lett 1998, 425:401-406. Eperon IC, Makarova OV, Mayeda A, Munroe SH, Caceres JF, Hayward DG, Krainer AR: Selection of alternative 5' splice sites: role of U1 snRNP and models for the antagonistic effects of SF2/ASF and hnRNP A1. Mol Cell Biol 2000, 20:8303-8318. Hieronymus H, Silver PA: A systems view of mRNP biology. Genes Dev 2004, 18:2845-2860. Mody M, Cao Y, Cui Z, Tay KY, Shyong A, Shimizu E, Pham K, Schultz P, Welsh D, Tsien JZ: Genome-wide gene expression profiles of the developing mouse hippocampus. Proc Natl Acad Sci U S A 2001, 98:8862-8867. Mills JC, Andersson N, Hong CV, Stappenbeck TS, Gordon JI: Molecular characterization of mouse gastric epithelial progenitor cells. Proc Natl Acad Sci U S A 2002, 99:14819-14824. Stappenbeck TS, Hooper LV, Gordon JI: Developmental regulation of intestinal angiogenesis by indigenous microbes via Paneth cells. Proc Natl Acad Sci U S A 2002, 99:15451-15455. Harris MN, Ozpolat B, Abdi F, Gu S, Legler A, Mawuenyega KG, Tirado-Gomez M, Lopez-Berestein G, Chen X: Comparative proteomic analysis of all-trans-retinoic acid treatment reveals systematic posttranscriptional control mechanisms in acute promyelocytic leukemia. Blood 2004, 104:1314-1323.

----------Hasta aquí la conversión del XPDF.----------

BioMed Central

BMC Developmental Biology

Open Access Research article

A genomewide in situ hybridization map of RNAbinding proteins reveals anatomically restricted expression in the developing mouse brain

Adrienne E McKee†1,2, Emmanuel Minet†2,3, Charlene Stern2, Shervin Riahi2, Charles D Stiles2,4 and Pamela A Silver*1,2

Address: 1Department of Systems Biology, Harvard Medical School, Boston, MA 02115 USA, 2Department of Cancer Biology, The DanaFarber Cancer Institute, Boston, MA 02115 USA, 3URBCFUNDP, 61 rue de Bruxelles, 5000 Namur, Belgium and 4Department of Microbiology and Molecular Genetics, Harvard Medical School, Boston, MA 02115 USA

Email: Adrienne E McKee adrienne_mckee@student.hms.harvard.edu; Emmanuel Minet emmanuel.minet@fundp.ac.be; Charlene Stern cstern@foleyhoag.com; Shervin Riahi shervin@gmail.com; Charles D Stiles charles_stiles@dfci.harvard.edu; Pamela A Silver* pamela_silver@dfci.harvard.edu

* Corresponding author †Equal contributors

Abstract

Background: In eukaryotic cells, RNAbinding proteins (RBPs) contribute to gene expression by regulating the form, abundance, and stability of both coding and noncoding RNA. In the vertebrate brain, RBPs account for many distinctive features of RNA processing such as activitydependent transcript localization and loca lized protein synthesis. Several RBPs with activities that are important for the proper function of adult brain have been identified, but how many RBPs exist and where these genes are expressed in the developing brain is uncharacterized.

Results: Here we describe a comprehensive catalogue of the unique RBPs encoded in the mouse genome and provide an online database of RBP expression in developing brain. We identified 380 putative RBPs in the mouse genome. Using in situ hybridization, we visualized the expression of 323 of these RBP genes in the brains of developing mice at embryonic day 13.5, when critical fate choice decisions are made and at P0, when major structural components of the adult brain are apparent. We demonstrate i) that 16 of the 323 RBPs examined show neuralspecific expression at the stages we examined, and ii) that a far larger subset (221) shows regionally restricted expression in the

brain. Of the regionally restricted RBPs, we describe one group that is preferentially expressed in the E13.5 ventricular areas and a second group that shows spatially restricted expression in post mitotic regions of the embryonic brain. Additionally, we find a subset of RBPs that share the same complex pattern of expression, in proliferating regions of the embryonic and postnatal NS and peripheral tissues.

Conclusion: Our data show that, in contrast to their proposed ubiquitous involvement in gene regulation, most RBPs are not uniformly expres sed. Here we demonstrate the regionspecific expression of RBPs in proliferating vs. postmitotic brain regions as well as celltypespecific RBP expression. We identify uncharacterized RBPs that exhibit neuralspecific expression as well as novel RBPs that show expression in nonneural tissues. The data presented here and in an online database provide a visual filter for the functional analysis of individual RBPs.

Published: 20 July 2005

BMC Developmental Biology 2005, 5:14 doi:10.1186/1471213X514

Received: 06 May 2005

Accepted: 20 July 2005

This article is available from: http://www.biomedcentral.com/1471213X/5/14

Background

The ordered production and differentiation of cell types that occurs during nervous system (NS) development relies upon tightly regulated gene expression. In neural cells, spatial and temporal gene regulation occurs through both transcriptional and posttranscriptional mecha nisms. While the transcriptional networks that direct neu ral cell fate and govern cell shape, position, and connectivity have been well studied [13], the posttran

scriptional influences on neural development and gene expression are less well understood.

At the core of posttranscriptional gene regulation are RNAbinding proteins (RBPs). Proteins containing canon ical RNAbinding domains (RBDs) are involved in numer ous steps of nuclear and cytoplasmic RNA processing [4]. Through mRNA capping, splicing, editing, polyadenyla tion and nonsensemediated decay, RBPs modulate the diversity of transcribed genes [46]. RBPs also affect the processing of noncoding RNAs [7]. Specific RBPs addi tionally enable asymmetric RNA distribution and transla tional regulation [810], two phenomena that are critical for affecting localized protein synthesis [11,12].

The importance of posttranscriptional processing in NS gene regulation is underscored by functional examples of specific RBPs [13,14]. For instance, the neuronalspecific factor Nova1 regulates splicing of premRNAs that encode components of inhibitory synapses [15]. Mice lacking Nova1 die postnatally due to aberrant regulation of apoptotic neuronal death [16]. As a second example, RBPs encoded by the quaking and Musashi loci promote glial cell fate [17] and CNS stem cell selfrenewal [18] by stabilizing transcripts involved in cell differentiation. Thirdly, the fragile X mental retardation protein, members of the ELAV/Hu protein family, and the Staufen proteins are involved in targeting and translational regulation of dendritic transcripts [1921]. Additionally, the finding that longterm memory requires de novo protein synthesis highlights the significance of posttranscriptional proc esses in neural function [22,23].

Despite our knowledge of several key RBPs, much of the understanding of RBPs in the brain comes from studies of adult animals or neural cell lines. Thus, how the func tional class of RBPs contributes to the positioning, growth, and diversification of cells in the developing brain is not well understood. One step towards increasing our understanding RBPs is to resolve where they are

expressed. Here, we utilize the approach of in situ hybrid ization mapping [2426] to investigate the expression of 323 RBPs within the developing mouse brain. Two stages of development were characterized, embryonic day 13.5 (E13.5), when critical cellular fate choice decisions are made and postnatal day 0 (P0), when the major structural

components of the brain are apparent. We find that, in contrast to their proposed ubiquitous involvement in gene regulation, most RBPs are not uniformly expressed. The majority of RBPs profiled demonstrates spatially restricted expression in the brain or in other peripheral tis sues examined. The data presented here and in an online database afford a visual filter for the functional analysis of individual RBPs in the developing mammalian NS.

Results

Mouse RBPs were identified according to gene sequence The RNA recognition motif (RRM), the hnRNP Khomol ogy (KH) domain, and the doublestranded RNAbinding domain (dsRM) are evolutionarily conserved, wellchar acterized domains known to bind either single or double stranded RNA [2729]. Sequence similarity searches and structural analyses of these domains have led to the ability to predict other RBPs based on primary coding sequence [29]. To identify unique genomic loci that encode puta tive RBPs in the mouse genome, we analyzed existing pub lic [30,31] and private [32] databases for sequences containing one or more RBD. Candidates were classified as RBPs only if their predicted protein sequence contained a Protein Families Database (Pfam)defined RBD [31]. We identified 290 genes harboring one or more RRM, KH, or dsRM sequences. We also identified 32 genes encoding other domains shown to interact with RNA, including the zinc knuckle, Gpatch, PIWI, DEAD box RNA helicase, and TUDOR domains. Finally, as the absence of a canon ical RBD does not preclude interaction with RNA, we

sought 58 additional genes known or predicted to be asso ciated with RNA processing. In total, this collection con tains 380 putative RBPs. Additional file 1 lists the number of genes, per RBD, identified and analyzed by in situ hybridization. A list of all genes and primer sequences is given in Additional file 2.

RBP expression in the developing mouse brain was analyzed by in situ hybridization

To localize RBP expression, we preformed in situ hybridi zation on whole head tissue sections of E13.5 embryos and P0 mice. We designed genespecific primers to pro duce 400–700 bp probes for 340 candidate RBPs. These primer sets were used to perform PCR on cDNA prepared from embryonic or P0 mouse brains. A small number of probes were obtained from mouse intestine, liver, kidney, or testes cDNA. 323 genes (95%) showed positive PCR products (data not shown). Following subcloning, anti sense digoxygeninlabeled riboprobes were prepared and hybridized against coronal head and transverse upper body sections (to include the brain and spinal cord, respectively). Digital images of the entire in situ hybridiza tion set have been deposited in the Mahoney RNABind ing Protein Expression Database [33].

= Page 2 =

RBPs exhibit restricted expression in the developing mouse brain

Several neuralspecific RBPs have been identified, yet how many others demonstrate this degree of specificity is unknown. Of the genes examined we found 16 RBPs (listed in Additional file 2) that exhibit NSrestricted

expression in the tissues analyzed. Among this list are known examples of neuronalspecific RBPs including Nova1 [34], the ELAV/Hu proteins B, C, and D [35], and Ataxin 2 binding protein 1 (A2bp1) [36] but additionally include putative RBPs for which expression has not been reported. With the exception of one gene that was only detected at E13.5, all (15/16) of these RBPs appear brain or NSspecific at both developmental stages in the tissues analyzed. Overall, these RBP encoding genes are not lim ited in expression to one brain region but are found in multiple brain or NS structures.

RBPs show spatially restricted expression in anatomically distinct brain regions

We find that greater than half of the RBPs profiled exhibit spatially restricted expression. Of the 323 genes exam ined, 221 demonstrate localized, enriched expression in one or more discrete brain regions in addition to detecta ble expression in nonNS tissues. We divided the E13.5 and P0 CNS into five and eight general areas for annota tion, respectively: the E13.5 precortical area, the striatum (and other basal ganglia), the periventrical areas, hind brain, and spinal cord, as well as the P0 cortex, striatum, hippocampus, thalamus, hypothalamus, midbrain, hind brain, and spinal cord. The presence or absence of expres sion for each RBP was analyzed visually at each location and is annotated in Additional file 3. Very few of the 221

RBPs with spatially restricted expression patterns were expressed in only one brain region, however most (73%) showed restricted expression at both developmental stages (Additional file 3).

We observe multiple RBPs that demonstrate regionspe cific expression in the E13.5 ventricular areas. Shown in Figure 1 are representative RBP genes that are transcribed in mitoticallyactive cells in the neuroepithelia of the developing telencephalon. Among the RBPs expressed in

this region occupied by neural progenitor cells, we find examples of mRNA export factors in addition to putative splicing factors and transcriptional regulators (Fig. 1). In all instances, expression in the embryonic lateral ventricu lar zone is accompanied by expression in the periventricu lar areas of the 3 rd and 4th E13.5 ventricles and often by heightened expression in the P0 subventricular zone [33]. Notably, we observed this pattern of expression for the dsRMcontaining Musashi proteins [33]. Our results are consistent with the documented expression of Msi1 and Msi2 [37,38].

Multiple RBPs show restricted expression in postmitotic regions of embryonic brain. Presented in Figure 2 are examples of four putative RBPs that demonstrate region specific expression in areas containing postmitotic neu rons. Transcripts of the genes encoding the RRM protein Brunol6 and the predicted zincknuckle protein 1500031H04Rik appear panneuronal at both develop mental stages (Fig. 2A, 2B and [33]). Expression of the RRMcontaining RIKEN gene 4930565A21 is most pro nounced in the ventral telencephalon, while D11Bwg0517e is found in the precortical layer, the

RBP expression in proliferative zones of the E13.5 mouse forebrainFigure 1

RBP expression in proliferative zones of the E13.5 mouse forebrain. In situ hybridization patterns for four RBPs on sections through the forebrain of E13.5 mice. Labels indicate Locuslink gene names. All images show the same magnification.

BMC Developmental Biology 2005, 5:14 http://www.biomedcentral.com/1471213X/5/14

thalamic area and hindbrain (Fig. 2C, 2D and [33]).

Among the genes that occupy postmitotic regions of the developing brain we additionally observe members of the ELAV/Hu family as well as other RBPs that have welldoc

umented neuronal expression [34,35].

RBPs demonstrate celltype specific expression in the P0 mouse retina

As our in situ hybridization analyses were performed on sections through whole head, we were able to visualize RBP expression in the developing retina. The vertebrate retina provides a distinctive system for studying CNS development as its seven major neural cell types are read ily distinguished from one another by their morphology and laminar position [39]. Shown in Figure 3 are exam ples of the diversity of RBP expression in the P0 retina. The RRMcontaining A2bp1 is expressed in the retinal gan glion cell layer (GCL), which contains primarily retinal ganglion cells and a small number of displaced amacrine cells (Fig 3A, 3B). The KHdomain encoding gene poly(rC) binding protein 3 (Pcbp3) shows dramatically enriched expression in the inner nuclear layer (INL) (Fig. 3C and 3D), possibly indicating localization to the bipo lar neuron cell bodies that occupy the scleral portion of the INL. Notably, both A2bp1 and Pcbp3 show restricted expression in postmitotic regions of the E13.5 and P0 brain [24,36]. Transcripts of the RRMencoding scaffold attachment factor B (Safb) and of the threeRRM contain ing SPOC gene Rbm15 are expressed in the outer neurob lastic layer of the retina (Fig. 3E–H). Safb, but not Rbm15, is additionally expressed in the GCL, possibly in the Müller glia. Both Safb and Rbm15 show enriched expres sion in neuroepithelia of the ventricular zone (Fig. 1 and [33]).

A systemsbased view of RBP expression

Gene regulation by RBPs is believed to occur through coordinated, combinatorial interactions with RNA. Dur ing the course of this study we identified multiple RBPs that are coordinately expressed in the brain and other tis

sues. We find 48 genes (listed in Additional file 4) that show elevated expression in proliferating areas of the embryonic and postnatal brain as well as in postnatal nasal epithelia, teeth, and thymus. Presented in Figure 4 are expression data for snRNP E and Son, two representa tive examples of this "synexpression group" of genes that share a similar, complex pattern of expression. Further examples are shown in Additional file 5. This same expres sion distribution has been observed for the polypyrimi dine tractbinding protein, PTBP1, and our data are consistent with previous findings [40]. Notably, the pro tein products of many of the genes listed are understood to interact either physically or genetically.

RBPs show restricted expression in nonNS tissues

As our analyses were performed on whole head and upper thoracic tissues, our data provide detailed information about RBP expression in developing cranial facial tissues. We identified putative RBPs that display tissuerestricted expression in nonNS structures (listed in Additional file 3). Figure 5 presents in situ hybridization results for two RRMencoding transcripts that show highly restricted expression in different epithelial tissues. The Riken gene 2210008M09 is transcribed in epithelia covering the facial skeleton (Fig. 5A, 5B), while the gene BC013481 is

RBP expression in postmitotic areas of the E13.5 mouse forebrainFigure 2

RBP expression in postmitotic areas of the E13.5 mouse forebrain. In situ hybridization patterns for four RBPs on sections through the forebrain of E13.5 mice. Labels indicate Locuslink gene names. bg, basal ganglia; hy, hypothalamus; nc, neo cortex. All images show the same magnification.

expressed in the choroid plexus (Fig. 5C) and in the lining of the intestine and placenta (Fig. 5D, 5E).

Discussion

Neural cells utilize multiple forms of posttranscriptional gene regulation. While RBPs are believed to be potent modulators of posttranscriptional processes, little is known about how this functional class is expressed in the developing brain. As a first step towards increasing our knowledge of RBPs we chose to investigate the spatial and temporal expression of genes that encode motifs known to interact with RNA. We find a small set of RBPs that show neuralspecific expression in the tissues analyzed.

Diversity of RBP expression in major cellular subtypes of the P0 retina

Figure 3

Diversity of RBP expression in major cellular sub types of the P0 retina. In situ hybridization for four repre sentative RBPs that exhibit laminarspecific expression in the P0 mouse retina. Labels indicate Locuslink gene names. A, B) A2bp1, C, D) Pcbp3, E, F) Safb, G, H) Rbm15. Panels A, C, E, and G show the same magnification. Panels B, D, F, and H show the same magnification. gcl, granule cell layer; inl, inner nuclear layer, onbl; outer neuroblastic layer.

Representative examples of RBP synexpression in E13.5 and P0 mouse tissues

Figure 4

Representative examples of RBP synexpression in E13.5 and P0 mouse tissues. snRNP E and Son are tran scribed in the perventricular areas of the E13.5 brain (A, E), in the P0 subventricular area of the lateral ventricle (B, F), in the external granule layer of the P0 cerebellum (C, G), as well as in postnatal developing teeth (D, H).

An even greater number of RBP genes however demon strate spatially restricted expression in distinct regions of the developing brain.

Within the CNS, most of the RBPs examined show non uniform, heightened expression in anatomically discrete structures. Tissue differences in the expression levels of individual genes could indicate distinctive protein requirements among cell types, beyond that of tissuespe cific RBPs [41]. There is precedent for differential require ments of individual RBPs, as tissuespecific RNA splicing is achieved partly through combinatorial, stoichiometric differences among splicing factors within various cells [42]. It is from this local enrichment within different cell types or tissues that we can begin to hypothesize as to the functional significance of individual genes as well as to the importance of groups of similarly expressed RBPs.

Our study has identified RBPs that display spatially restricted expression in distinct regions of the developing mouse brain. One set of RBPs (Fig. 1) is found in the E13.5 ventricular areas. A second set demonstrates spa tially restricted expression in postmitotic regions of E13.5 brain (Fig. 2). Based on their pattern of expression, these RBPs may have roles in neural proliferation, cell fate choice and cell migration, or in neuronal function, respec tively. We also identified novel RBPs that are expressed in tissues of mesodermal and endodermal origin (Fig. 5). The highly restricted expression of these genes may indi cate an explicit role for these RBPs in their respective epi thelia. Additionally, the celltype specificity RBPs found in the P0 retina (Fig. 3) illustrates the diversity of RBP expres sion. The specialized expression of these RBPs may be indicative of a dedicated function in the specified tissues.

By visual inspection of in situ hybridization data, we find

a subset of RBPs that are coordinately expressed in multi ple tissue types. These genes display heightened expres sion in the periventricular areas of the E13.5 brain and spinal cord as well as marked expression in the external granule layer of the P0 cerebellum, the lateral subventricu lar zones, and in teeth, nasal epithelia, and thymus (Fig. 4, Additional file 5, [33]). While not excluded from post mitotic tissues, these RBPs are predominately expressed in structures that are undergoing cell division.

Notably, the term 'synexpression group' has been used to describe collections of genes that function in a common process and share a similar complex spatial expression pattern in multiple tissues [43]. Among the synexpression group identified here we find examples of RBPs that are known to interact either physically or genetically (Addi tional file 4). For example, PTBP1 binds the splicing factors PSF [44] and hnRNP L [45] while SF2/ASF and hnRNP A1 select for 5' exon or exclusion or inclusion, respectively [46]. Our data provide visual support to a growing body of evidence that functionallyrelated tran scripts are posttranscriptionally coregulated [47].

Although the significance of certain splicing and mRNA export factor enrichment in proliferating regions is not known, data from multiple studies point to a role for RBPs in cell proliferation. During hippocampal development expression levels of RBPs were found to be high and then to dramatically decrease, as neurons transition from a pro liferating to a postmitotic state [48]. A number of RBPs were also identified as highly expressed in a molecular characterization of gastric epithelial progenitor cells [49,50]. Furthermore, protein levels of hnRNPs and snRNPs were found to be downregulated upon stimu lated growth inhibition of myeloid cells [51]. Therefore, it is likely that a role for RBPs during cell proliferation and cell fate determination exists in multiple tissue types.

Conclusion

In summary, the data presented here provide new insight into how a distinct functional gene class is expressed in the developing NS. We find that RBPs demonstrate

In situ hybridization profiling uncovers the nonneural, restricted expression of novel RBPs

Figure 5

In situ hybridization profiling uncovers the nonneu ral, restricted expression of novel RBPs. Data from ISH performed on (A, C) coronal E13.5 and on (B, D, E) E15 sag ittal sections are presented for RRMencoding RBPS. A, B) The Riken gene 2210008M09 is transcribed in epithelia cov ering the facial skeleton. CE) BC013481 is detected in the choroid plexus, in the intestinal lining, and in the lining of the placenta. Panels CE show the same magnification.

BMC Developmental Biology 2005, 5:14 http://www.biomedcentral.com/1471213X/5/14

regionspecific as well as celltype specific expression. In addition, we find that specific, proliferating regions of the embryonic and postnatal NS and peripheral tissues are similar in the expression of certain RBPs. These data serve as a starting point for functional investigations into the roles of RBPs in neural development and physiology.

Methods

In silico RBP identification

Putative RBP gene sequences were identified by homol ogybased whole genome screening using public and pri vate databases: Celera Panther Families, Protein Families Database (Pfam), and Genbank [3032]. Classification as an RBP was based on the presence of one or more RRM, KH, or dsRMs, as defined by Pfam databases [31]. Data bases were also mined for zincknuckle, Gpatch, PIWI,

DEADbox helicase and Tudor domaincontaining sequences and for known factors involved in mRNA splic ing, editing, transport, and stability. Genes with multiple RNAbinding domains were assigned to a single sub family. Unique gene identity was verified by LocusID numbers. As of March 1, 2004, a total of 357 unique genes were identified from these sources. An additional 26 RRM, KH, and dsRM proteins have been identified as of March 7, 2005.

PCR primer design

PCR primer pairs were designed for each identified RNA binding protein locus. PCR primer sequences were designed with approximately 60% GC content, spanning 400–700 base pairs of primarily the gene's coding sequence. Additional primer pairs were designed for tar gets that did not initially yield PCR products.

Cloning

Total RNA was obtained from E13.5, P0, or adult C57/BL6 mouse brains (Charles River Laboratories) by Trizol extraction (Invitrogen). Reverse transcription was per formed using Superscript II reverse transcriptase and oligodT (Invitrogen). PCR was performed with cDNA templates using 40 cycles, 60–65°C annealing tempera ture, and Platinum Taq (Invitrogen) as polymerase. For a few genes, PCR was performed with cDNA templates pre pared from adult brain, kidney, gut, liver, or testis tissues. Positive PCR products were cloned into TA cloning vectors (Invitrogen) and verified by restriction digest or DNA sequencing.

Probe synthesis

Gene fragments from verified plasmids were amplified by PCR using plasmid specific primers. Digoxigeninlabeled RNA probes were made, using PCR products as template and T7 or SP6 RNA polymerases (Roche). cRNA probes

were ethanol precipitated and quantified by spectrophotometry.

Tissue preparation

E13.5 embryos were directly fixed overnight in 4% para formaldehyde (0.1M PBS). P0 mice were transcardially perfused with 4% paraformaldehyde (0.1M PBS) and postfixed overnight at 4°C. After fixation, embryos and P0 mice were transferred to 20% sucrose overnight. The head, neck, and trunk were embedded separately in OCT (Tis sueTek) on dry ice and stored at 80°C. Serial cryostat sec tions (14 µm) were cut and mounted on Superfrost Plus slides (Fisher). Ten and twenty adjacent sets of sections were prepared from E13.5 embryos and P0 mice, respec tively, and were stored at 20°C until use.

Section in situ hybridization

In situ hybridization was performed according to Gray et al. [25]. Following pretreatment (Proteinase K), slides were prehybridized for 1h at 65°C in hybridization solu tion (50% formamide (Ambion), 5X SSC, 0.3 mg/ml yeast tRNA (Sigma), 100 µg/ml heparin (Sigma), 1X Denhardt's (Sigma), 0.1% tween, 5 mM EDTA). P0 and E13.5 brain sections were hybridized overnight with labeled RNA probe(0.8–1.2 µg/ml) at 65°C, washed in 2X SSC at 67°C, incubated with RNase A (1 µg/ml, 2X SSC) at 37°C, washed in 0.2X SSC at 65°C, blocked in PBS with 10% lamb sera, and incubated in alkaline phosphatase labeled antiDIG antibody (Roche) (1:2000, 10% sera) overnight. Sections were washed and color was visualized using NBT and BCIP in alkaline phosphatase buffer (100 mM Tris pH 9.5, 50 mM MgCl2, 100 mM NaCl, 0.1% tween20) con taining 75 µg/ml NBT (BioRad), 600 µg/ml BCIP (Roche). Staining was stopped after visual inspection. Sections were washed, fixed in 4% paraformaldehyde, and cover slipped in glycerol [25].

Image acquisition and RBP expression database

Images were acquired and analyzed as described [25]. Images were either scanned using a Nikon Coolscan 8000 slide scanner (4000 DPI) or digitally acquired using a Leica digital camera. Image levels have been modified in Photoshop (Adobe) for clarity. Full resolution scanned images were compressed using JPEG compression, quality 10, and have been deposited in the Mahoney RNABind ing Protein Expression Database [33].

Authors' contributions

AEM prepared tissue samples, performed data analysis and drafted the manuscript. EM performed data analysis and both EM and SR generated reagents, tissue samples, digitized the raw data, and helped build the website. CS contributed to the design of the study and prepared tissue samples. CDS and PAS conceived of the study, partici pated in its design and coordination and helped prepare the manuscript. All authors read and approved of the manuscript.

Additional material

Acknowledgements

We are grateful to Drs. Qiufu Ma and John Alberta for critical review of this manuscript and for assistance in this work. We thank Eric Tsung, Zhao hui Cai, and Matthew McCormack for designing the website. This work has been supported by the Bernard A. and Wendy J. Goldhirsh Foundation for Brain Tumor Research and by the Charles A. Dana Foundation. AEM is sup ported by an institutional training grant from the National Cancer Institute (T32CA09361). EM is funded as a FNRS Researcher through the Belgian National Research Fund and by the D. Collen Research Foundation VZW

and BAEF. CS received support from the American Cancer Society (PF02 12801MBC).

References

1. Ross SE, Greenberg ME, Stiles CD: Basic helixloophelix factors in cortical development. Neuron 2003, 39:1325.

2. Wilson SW, Houart C: Early steps in the development of the forebrain. Dev Cell 2004, 6:167181.

3. BallyCuif L, Hammerschmidt M: Induction and patterning of neuronal development, and its connection to cell cycle control. Curr Opin Neurobiol 2003, 13:1625.

4. Dreyfuss G, Kim VN, Kataoka N: MessengerRNAbinding pro teins and the messages they carry. Nat Rev Mol Cell Biol 2002, 3:195205.

5. Lasko P: Gene regulation at the RNA layer: RNA binding pro teins in intercellular signaling networks. Sci STKE 2003, 2003:RE6.

6. Orphanides G, Reinberg D: A unified theory of gene expression. Cell 2002, 108:439451.

7. Cullen BR: Transcription and processing of human microRNA precursors. Mol Cell 2004, 16:861865.

8. Huang YS, Carson JH, Barbarese E, Richter JD: Facilitation of den dritic mRNA transport by CPEB. Genes Dev 2003, 17:638653.

9. Antar LN, Bassell GJ: Sunrise at the synapse: the FMRP mRNP shaping the synaptic interface. Neuron 2003, 37:555558.

10. Tang SJ, Meulemans D, Vazquez L, Colaco N, Schuman E: A role for a rat homolog of staufen in the transport of RNA to neuronal dendrites. Neuron 2001, 32:463475.

11. Martin KC: Local protein synthesis during axon guidance and synaptic plasticity. Curr Opin Neurobiol 2004, 14:305310.

12. Huang YS, Richter JD: Regulation of local mRNA translation. Curr Opin Cell Biol 2004, 16:308313.

13. Agnes F, Perron M: RNAbinding proteins and neural develop ment: a matter of targets and complexes. Neuroreport 2004, 15:25672570.

14. PerroneBizzozero N, Bolognani F: Role of HuD and other RNA binding proteins in neural development and plasticity. J Neu

rosci Res 2002, 68:121126.

15. Ule J, Jensen KB, Ruggiu M, Mele A, Ule A, Darnell RB: CLIP identi

fies Novaregulated RNA networks in the brain. Science 2003,

302:12121215.

16. Jensen KB, Dredge BK, Stefani G, Zhong R, Buckanovich RJ, Okano

HJ, Yang YY, Darnell RB: Nova1 regulates neuronspecific

Additional File 1

RNAbinding proteins identified in silico and profiled by in situ hybridization. List of annotated RNAbinding domains and the number of family members that were identified in silico and analyzed by in situ hybridization.

Click here for file

[http://www.biomedcentral.com/content/supplementary/1471

213X514S1.xls]

Additional File 2

List of 380 genes identified as putative RBPs in the mouse genome and analyzed in this study. Columns indicate LocusID, gene name, type of RBD, primer sequences used to isolate the target cDNA, the size of the cDNA fragment, the presence call by PCR from E13.5 and P0 brain cDNA, cloning status ('c' indicates cloned, 'u' indicates uncloned, 'small' indicates that the target gene had less than 400 bp of unique sequence, 'na' indicates that cloning was not attempted), the RNA polymerase used to generate the antisense riboprobe, the tissue from which the cDNA was isolated (if not from E13.5 or P0 mouse brain), and whether the gene was analyzed by in situ hybridization ('x' indicates yes).

Click here for file

[http://www.biomedcentral.com/content/supplementary/1471

213X514S2.xls]

Additional File 3

Complete list of gene expression patterns for all in situ hybridizations performed. Of the 323 RBPs examined, 221 showed restricted expression patterns in the brain. The remaining genes either show restricted expres sion in nonneural tissues, ubiquitous expression that is difficult to distin guish from background, or no expression. Caution is needed in

interpreting the results. First, nonexpression could be due to the sensitivity limit of nonradioactive in situ hybridization. Second, the background level of individual probes may differ. Third, some probes with high back ground hybridization may mask the real expression of the transcript. Fourth, we cannot rule out the possibility that some probes may show var iable levels of background hybridization in different brain areas, resulting in a false positive signal. Columns AD describe the LocusID, gene name, type of RBD, and number (internal Mahoney reference number). Col umns E and, L (E13.5, P0 "Informativity"): "1" for restricted expression in the nervous system and "0" for either ubiquitous expression that is dif ficult to distinguish from background or no expression. As noted in Gray et al [25], some of the genes in the "0" category show uneven signals in different brain regions and are also annotated in the subsequent columns. Columns F and M (E13.5, P0 "Specificity"): "1" for restricted expression in neural tissues only, "2" for restricted expression in neural tissue with distinguishable expression in nonneural tissue, "3" for ubiquitous or no expression, and "4" for expression in nonneural tissues only. Columns G K and NU (E13.5, P0 "Expression"): "2" for expression, "1" for ubiq uitous expression or background, "0" for no expression.

Click here for file

[http://www.biomedcentral.com/content/supplementary/1471 213X514S3.xls]

Additional File 4

RNAbinding proteins belonging to a synexpression group. Complete list of RBPs that demonstrate a similar complex pattern of expression. Col umns AD describe the LocusID, gene name, type of RBD, and number (internal Mahoney reference number).

Click here for file

[http://www.biomedcentral.com/content/supplementary/1471 213X514S4.xls]

Additional File 5

Examples of RBP synexpression in E13.5 and P0 mouse tissues. Addi tional examples of RBPs that share a similar pattern of expression. Shown are in situ hybridization results of expression in the periventricular areas of the E13.5 brain (A, E, I, M, Q), in the subventricular area of the P0

lateral ventricle (B, F, J, N, R), in the external granule layer of the P0 cerebellum (C, G, K, O, S), as well as in postnatal developing teeth (D, H, L P, T). AD) Refbp1, EH) hnRNP A1, IL) PTBP1, MP) Sfpq, Q R) Hnrpl. Panels A, B, E, F, I, J, M, N, Q, R show the same magnification. Panels C, D, G, H, K, L, O, P, S, T show the same magnification.

Click here for file

[http://www.biomedcentral.com/content/supplementary/1471213X514S5.png]

alternative splicing and is e ssential for neuronal viability.

Neuron 2000, 25:359371.

17. Larocque D, Galarneau A, Liu HN, Scott M, Almazan G, Richard S:

Protection of p27(Kip1) mRNA by quaking RNA binding pro

teins promotes oligodendrocyte differentiation. Nat Neurosci

2005, 8:2733.

18. Sakakibara S, Nakamura Y, Yoshida T, Shibata S, Koike M, Takano H,

Ueda S, Uchiyama Y, Noda T, Okano H: RNAbinding protein

Musashi family: roles for CNS stem cells and a subpopulation

of ependymal cells revealed by targeted disruption and anti

sense ablation. Proc Natl Acad Sci U S A 2002, 99:1519415199.

19. Pascale A, Gusev PA, Amadio M, Dottorini T, Govoni S, Alkon DL,

Quattrone A: Increase of the RNAbinding protein HuD and

posttranscriptional upregulation of the GAP43 gene during

spatial memory. Proc Natl Acad Sci U S A 2004, 101:12171222.

20. Jin P, Alisch RS, Warren ST: RNA and microRNAs in fragile X

mental retardation. Nat Cell Biol 2004, 6:10481053.

21. Dubnau J, Chiang AS, Grady L, Barditch J, Gossweiler S, McNeil J,

Smith P, Buldoc F, Scott R, Certa U, Broger C, Tully T: The staufen/

pumilio pathway is involved in Drosophila longterm

memory. Curr Biol 2003, 13:286296.

22. Miller S, Yasuda M, Coats JK, Jones Y, Martone ME, Mayford M: Dis

ruption of dendritic translation of CaMKIIalpha impairs sta

bilization of synaptic plasticity and memory consolidation.

Neuron 2002, 36:507519.

23. Kang H, Schuman EM: A requirement for local protein synthesis

in neurotrophininduced hippocampal synaptic plasticity. Sci

ence 1996, 273:14021406.

24. Reymond A, Marigo V, Yaylaoglu MB, Leoni A, Ucla C, Scamuffa N,

Caccioppoli C, Dermitzakis ET, Lyle R, Banfi S, Eichele G, Antonarakis

SE, Ballabio A: Human chromosome 21 gene expression atlas

in the mouse. Nature 2002, 420:582586.

25. Gray PA, Fu H, Luo P, Zhao Q, Yu J, Ferrari A, Tenzen T, Yuk DI,

Tsung EF, Cai Z, Alberta JA, Cheng LP, Liu Y, Stenman JM, Valerius

MT, Billings N, Kim HA, Greenberg ME, McMahon AP, Rowitch DH,

Stiles CD, Ma Q: Mouse brain organization revealed through

direct genomescale TF expression analysis. Science 2004,

306:22552257.

26. Gitton Y, Dahmane N, Baik S, Ruiz i Altaba A, Neidhardt L, Scholze M, Herrmann BG, Kahlem P, Benkahla A, Schrinner S, Yildirimman R, Herwig R, Lehrach H, Yaspo ML: A gene expression map of human chromosome 21 orthologues in the mouse. Nature 2002, 420:586590.

27. Saunders LR, Barber GN: The dsRNA binding protein family: critical roles, dive rse cellular functions. Faseb J 2003, 17:961983.

28. Nagai K: RNAprotein complexes. Curr Opin Struct Biol 1996, 6:5361.

29. Burd CG, Dreyfuss G: Conserved structures and diversity of functions of RNAbinding proteins. Science 1994, 265:615621.

30. Wheeler DL, Church DM, Edgar R, Federhen S, Helmberg W, Mad den TL, Pontius JU, Schuler GD, Schriml LM, Sequeira E, Suzek TO, Tatusova TA, Wagner L: Database resources of the National Center for Biotechnology Information: update. Nucleic Acids Res 2004, 32:D3540.

31. Sonnhammer EL, Eddy SR, Birney E, Bateman A, Durbin R: Pfam: multiple sequence alignments and HMMprofiles of protein domains. Nucleic Acids Res 1998, 26:320322.

32. Thomas PD, Campbell MJ, Kejariwal A, Mi H, Karlak B, Daverman R, Diemer K, Muruganujan A, Narechania A: PANTHER: a library of protein families and subfamilies indexed by function. Genome Res 2003, 13:21292141.

33. http://mahoney.chip.org/mahoney/RBP. .

34. Buckanovich RJ, Posner JB, Darnell RB: Nova, the paraneoplastic Ri antigen, is homologous to an RNAbinding protein and is specifically expressed in the developing motor system. Neu ron 1993, 11:657672.

35. Okano HJ, Darnell RB: A hierarchy of Hu RNA binding proteins in developing and adult neurons. J Neurosci 1997, 17:30243037.

36. Kiehl TR, Shibata H, Vo T, Huynh DP, Pulst SM: Identification and expression of a mouse ortholog of A2BP1. Mamm Genome 2001, 12:595601.

37. Sakakibara S, Okano H: Expression of neural RNAbinding pro teins in the postnatal CNS: implications of their roles in neu

ronal and glial cell development. J Neurosci 1997, 17:83008312.

38. Sakakibara S, Nakamura Y, Satoh H, Okano H: Rnabinding protein Musashi2: developmentally regulated expression in neural precursor cells and subpopulations of neurons in mamma lian CNS. J Neurosci 2001, 21:80918107.

39. Blackshaw S, Harpavat S, Trimarchi J, Cai L, Huang H, Kuo WP, Weber G, Lee K, Fraioli RE, Cho SH, Yung R, Asch E, OhnoMachado L, Wong WH, Cepko CL: Genomic analysis of mouse retinal development. PLoS Biol 2004, 2:E247.

40. Lillevali K, Kulla A, Ord T: Comparative expression analysis of the genes encoding polypyrimidine tract binding protein (PTB) and its neural homologue (brPTB) in prenatal and postnatal mouse brain. Mech Dev 2001, 101:217220.

41. Zhang W, Liu H, Han K, Grabowski PJ: Regionspecific alternative splicing in the nervous system: implications for regulation by the RNAbinding protein NAPOR. Rna 2002, 8:671685.

42. Grabowski PJ, Black DL: Alternative RNA splicing in the nerv ous system. Prog Neurobiol 2001, 65:289308.

43. Niehrs C, Pollet N: Synexpression groups in eukaryotes. Nature 1999, 402:483487.

44. Patton JG, Porro EB, Galceran J, Tempst P, NadalGinard B: Cloning and characterization of PSF, a novel premRNA splicing factor. Genes Dev 1993, 7:393406.

45. Hahm B, Cho OH, Kim JE, Kim YK, Kim JH, Oh YL, Jang SK: Polypy rimidine tractbinding protein interacts with HnRNP L. FEBS Lett 1998, 425:401406.

46. Eperon IC, Makarova OV, Mayeda A, Munroe SH, Caceres JF, Hay ward DG, Krainer AR: Selection of alternative 5' splice sites: role of U1 snRNP and models for the antagonistic effects of SF2/ASF and hnRNP A1. Mol Cell Biol 2000, 20:83038318.

47. Hieronymus H, Silver PA: A systems view of mRNP biology. Genes Dev 2004, 18:28452860.

48. Mody M, Cao Y, Cui Z, Tay KY, Shyong A, Shimizu E, Pham K, Schultz P, Welsh D, Tsien JZ: Genomewide gene expression profiles of the developing mouse hippocampus. Proc Natl Acad Sci U S A 2001, 98:88628867.

49. Mills JC, Andersson N, Hong CV, Stappenbeck TS, Gordon JI: Molec
ular characterization of mouse gastric epithelial progenitor
cells. Proc Natl Acad Sci U S A 2002, 99:1481914824.
50. Stappenbeck TS, Hooper LV, Gordon JI: Developmental regula
tion of intestinal angiogenesis by indigenous microbes via
Paneth cells. Proc Natl Acad Sci U S A 2002, 99:1545115455.
51. Harris MN, Ozpolat B, Abdi F, Gu S, Legler A, Mawuenyega KG,
TiradoGomez M, LopezBerestein G, Chen X: Comparative pro
teomic analysis of alltransretinoic acid treatment reveals
systematic posttranscriptional control mechanisms in acute
promyelocytic leukemia. Blood 2004, 104:13141323.

Comentarios:

En la conversión del ejemplo anterior podemos ver las diferencias al utilizar un conversor u otro. Dependiendo de cual sea presenta ciertas ventajas o inconvenientes que pasamos a analizar en detalle comparando uno a uno.

En una primera impresión, vemos como el A-PDF no entiende correctamente los fines de linea y a pesar de no mantener el “layout” del archivo original, mantiene la longitud de la frase que ocupaba en la página físicamente. En el ejemplo con el XPDF se ve como si entiende sin problemas el fin de linea y muestra el texto de forma continuada.

Otra característica no deseada en la conversión es que siempre se muestra el pie de página, indistintamente del programa escogido. Información esta, completamente prescindible para el procesado posterior de la infromación. Esta característica se muestra en todos los programas analizados.

Asi mismo, en todos los programas vemos como los pie de foto muestran dos veces el texto que está en negrita. Independientemente del programa escogido.

Por último, y ya al final del documento científico, nos encontramos con la bibliografía. Aquí encontramos serios problemas con el XPDF, que no entiende los índices de las referencias, de modo que inserta los números de los índices de una misma página todos seguidos. Y a continuación el texto de cada referencia, de modo que es prácticamente ilegible.